Supplemental Materials to "Robust Variable and Interaction Selection for Logistic Regression and General Index Models"

Yang Li, Jun S. Liu
Department of Statistics, Harvard University

Contents

1 Additional empirical study results

1.1 Simulation study on logistic regression with only main effects

Here we report a few more simulation examples, in which there are only main effects by no interactions, to compare SODA with Lasso (denoted as Lasso-Logistic) on logistic regression variable selection. Examples 0.1 and 0.2 illustrate two simulation settings, respectively. In both examples, we simulated predictors $\mathbf{X}$ from the multivariate normal distribution with covariance matrix $\mathbf{C}$. In Example 0.1 we set $\mathbf{C}$ to have power decay correlations between variables, and in Example 0.2 we obtained $\mathbf{C}$ from a real dataset. Let $\mathbf{Q}$ denote the Fisher information matrix of the form,

$$\mathbf{Q} \equiv \mathbb{E}\left\{-\nabla^2 \log p\left(Y \mid \mathbf{X}, \boldsymbol{\theta}_0\right)\right\} = \frac{\exp\left(\boldsymbol{\theta}_0^T \mathbf{X}\right)}{\left(1+\exp\left(\boldsymbol{\theta}_0^T \mathbf{X}\right)\right)^2} \mathbf{X}\mathbf{X}^T. \tag{1}$$

Let sub-matrices $\mathbf{Q}_{21} = \mathbf{Q}_{\mathcal{P}^c,\mathcal{P}}$ and $\mathbf{Q}_{11} = \mathbf{Q}_{\mathcal{P},\mathcal{P}}$. The consistency of Lasso-Logistic requires the "incoherence" (Ravikumar et al. 2010) or the "irrepresentable" (Zhao and Yu 2006) condition that there exists an $\alpha \in (0, 1]$ such that

$$\left\|\mathbf{Q}_{21}\mathbf{Q}_{11}^{-1}\right\|_\infty \leq 1-\alpha.$$

Let $\boldsymbol{c} = \mathbf{Q}_{21}\mathbf{Q}_{11}^{-1}$, then $\boldsymbol{c}$ has the length as the number of unrelated predictors, and the incoherence condition requires each element of $\boldsymbol{c}$ to be smaller than 1.

For each setting, we randomly generated 100 datasets from the logistic regression model with $n = 100, 150, \ldots, 2000$ observations, and applied SODA and Lasso-Logistic to each data set. We used EBIC$_\gamma$ as criterion for both SODA and Lasso-Logistic. Lasso-Logistic fitted a solution path of selected predictors, and chose the optimal set of predictors with the lowest EBIC$_\gamma$. In simulation studies, we set $\gamma = 0.5$. We calculated the average number of false negatives (FN) and false positives (FP), and the percentage of correct fits (PCF), which is the percentage of times that the selected set is the true set $\mathcal{A}$. For SODA, any selected interaction term would also be considered as a false positive.

Example 0.1. Let $p = 1000$, and we randomly selected 5 true predictors with coefficients $\beta_{0,j} \sim \text{Unif}\,[0.5, 2]$, $j \in \mathcal{A}$. The covariance matrix is set to have power decay correlation such that $\mathbf{C}_{i,j} = (0.5)^{|i-j|}$. Following a similar argument as Corollary 3 of Zhao and Yu (2006), it is easy to show that $\mathbf{X}$ satisfies the incoherence condition. The histogram of elements of $\boldsymbol{c}$ in log-scale for one simulation run is plotted in Figure 1, and it is shown that no $c_j \geq 1$ in $\boldsymbol{c}$. As shown in Figure 2, SODA and Lasso-Logistic had very similar performances under this setting.

Example 0.2. In this example, we also randomly selected 5 true predictors with coefficients $\beta_{0,j} \sim \text{Unif}\,[0.5, 2]$, $j \in \mathcal{A}$. The covariance matrix $\mathbf{C}$ was set to be the sample covariance matrix of the Michigan lung cancer dataset (Beer et al. 2002) with $p = 5,217$ genes. So Example 0.2 had a much higher dimension than Example 0.1. The histogram of elements of $\boldsymbol{c}$ in log-scale for one simulation run is plotted in Figure 1.

As illustrated by Figure 1, in Example 0.2 many predictors are highly correlated with each other, and thus the incoherence condition is strongly violated. As shown in Figure 2, in this case Lasso-Logistic had a very poor performance whereas SODA performed robustly. As n increases, the Lasso-Logistic's total number of FPs and FNs increased, and PCF stayed at zero. In contrast, SODA had an increasing probability of selecting the correct model $\mathcal{A}$ as n increases. In both examples, SODA did not select any interaction term for all simulations.

1.2 Real data analysis on prostate cancer dataset

The microarray technology is widely used for measuring expression abundance of genes. There have been tremendous efforts on building classification methods to diagnose cancer patients from microarray data. In Singh et al. (2002), researchers measured the gene expressions of 52 prostate cancer patients and 50 controls on $p = 6,033$ genes. The goal is to predict whether a person has prostate cancer from the expression of those genes. Efron (2009) proposed an empirical Bayes approach for large-scale classification, and compared its performance with that of the shrunken centroids method proposed in Tibshirani et al. (2002). With different thresholds, the shrunken centroids method and the empirical Bayes report selected set of predictors truncated at different sizes. A common way of applying the two methods is to obtain selected predictors with different thresholds, and pick the best set by cross-validation (CV). We implemented these two methods

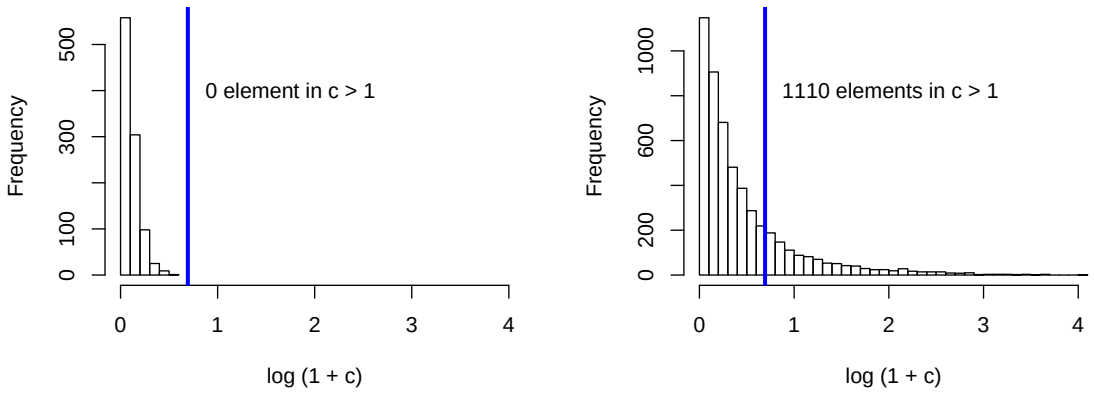

Figure 1: Histogram of elements of $\boldsymbol{c}$ in one simulation run for Example 0.1 (Left) and 0.2 (Right).

and calculated the CV prediction error of the selected gene set of different sizes. The number of selected genes and the 10-fold CV error rate (CVE) of two methods on different thresholds are shown in Table 1.

The shrunken centroids method selected 377 genes at the threshold $\lambda = 2.16$ that achieved the lowest CVE, and empirical Bayes method selected the best set with 51 genes. In the solution path of Lasso-Logistic, the lowest $\text{EBIC}_{0.5}$ was achieved at 133.3 with 3 genes, and the corresponding CVE was 17%. SODA selected 6 main effects and 0 interaction with the $\text{EBIC}_{0.5}$ score at 93.4 and the CVE at 6%.

MDR failed to converge on this dataset. In particular, MDR selected as many genes as possible until the number of selected genes was the same as the number of samples in the smaller class (50). Subsequently the estimated covariance matrix for the smaller class became singular and the procedures could not proceed. ΔBIC_G, defined as the difference of BIC_G two adjacent steps (see the main paper), is shown for each step in Table 1. MDR proceeds if $\Delta\text{BIC}_G < 0$ and eventually selects 49 genes with CVE 52%.

We applied IIS-SQDA to this problem by running the R code provided by its authors. But for this dataset IIS-SQDA did not finish the analysis in 48 hours. The reason is as noted in Fan et al. (2015) that IIS needs to estimate the precision matrices, which can be very slow when the number of predictors p is large.

It is worth noting that although the final model selected by SODA has only main effect terms with no interactions, SODA's model managed to outperform the one selected by Lasso-logistic in terms of both the $\text{EBIC}_{0.5}$ score and the CVE. This strong result obtained by SODA is also surprising to us, indicating that EBIC is a good criterion to follow and our stepwise approach is a better optimizer of EBIC than Lasso. Indeed, when one moves away from the L_1 regularization realm but adopts the L_0 regularization framework (such as AIC, BIC, EBIC, etc.), Lasso can no longer guarantee to find the optimal solution. We consistently observed that SODA outperformed Lasso in various simulation settings in finding configurations with a low $\text{EBIC}_{0.5}$ score.

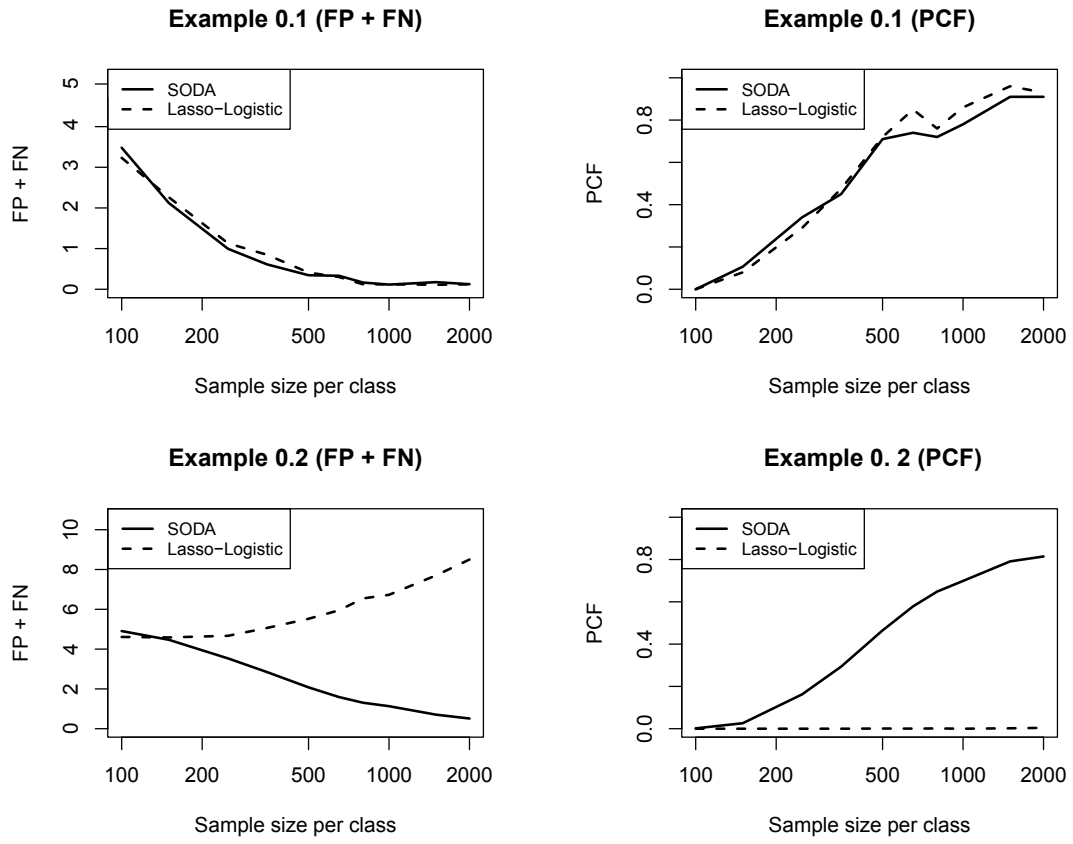

Figure 2: Simulation study results for Example 0.1 (top) and 0.2 (bottom). FP: average number of false positives. FN: average number of false negatives. PCF: percentage of times of selecting correct model $\mathcal{A}$.

2 Proofs

Let $\|\cdot\|_2$, $\|\cdot\|_\infty$, $\|\cdot\|_{\text{sp}}$ and $\|\cdot\|_{\text{F}}$ respectively denote L_2, L_∞, spectral and Frobenius norms. Let $\mathbb{I}\{\cdot\}$ denote the identify function such that it takes value 1 if the statement within $\{\cdot\}$ is true and takes value 0 otherwise. Suppose $\mathbf{A}$ and $\mathbf{B}$ are two square symmetric matrices with same dimensions, then $\mathbf{A} \succeq \mathbf{B}$ indicates matrix $\mathbf{A} - \mathbf{B}$ is positive-definite. Suppose $\mathcal{C}$ is a predictor set, then $\mathcal{C}^c$ denotes its complementary set, i.e. $\mathcal{C}^c = \{1, \ldots, p\} \setminus \mathcal{C}$.

Let $\boldsymbol{\theta}_\mathcal{S} = (\boldsymbol{\theta}_{1,\mathcal{S}}, \cdots, \boldsymbol{\theta}_{K,\mathcal{S}})$ denote the parameter vector where coefficients are set as 0 for terms not in $\mathcal{S}$, where $\boldsymbol{\theta}_{k,\mathcal{S}}$ denotes corresponding coefficients for class k. By definition $\boldsymbol{\theta}_{K,\mathcal{S}} \equiv \mathbf{0}$. The log-likelihood for $\boldsymbol{\theta}_\mathcal{S}$ on observations $\{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ is

$$l_n\left(\boldsymbol{\theta}_\mathcal{S}\right) = \sum_{i=1}^{n} \left\{ \boldsymbol{\theta}_{y_i,\mathcal{S}}^T \mathbf{z}_i - \log\left(1 + \sum_{l=1}^{K-1} \exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T \mathbf{z}_i\right)\right)\right\}.$$

Let $\mu_{i,k}\left(\boldsymbol{\theta}_\mathcal{S}\right)$ denote the estimated probability of ith observation being in class k with parameters

Shrunken centroid		**Empirical Bayes**		**MDR**			**Lasso-Logistic**			**SODA**		
#P	CVE	#P	CVE	ΔBIC	#P	CVE	$\text{EBIC}_{0.5}$	#P	CVE	$\text{EBIC}_{0.5}$	#M/#I	CVE
0	0.52	1	0.34	-68	1	0.28	140.3	1	0.28	140.3	1 / 0	0.28
1	0.48	5	0.30	-74	2	0.28	134.2	2	0.22	133.0	2 / 0	0.21
4	0.41	10	0.27	-68	5	0.17	**133.3**	**3**	**0.17**	124.1	3 / 0	0.17
35	0.30	15	0.26	-57	10	0.11	141.4	4	0.16	107.8	4 / 0	0.15
80	0.16	20	0.21	-64	15	0.11	144.3	5	0.16	100.9	5 / 0	0.09
172	0.10	25	0.20	-74	20	0.14	151.3	6	0.13	**93.4**	**6 / 0**	**0.06**
377	0.09	30	0.15	-90	25	0.20	156.7	7	0.13			
866	0.12	35	0.11	-108	30	0.28	151.9	8	0.11			
1,931	0.23	40	0.12	-122	35	0.31	158.4	9	0.11			
3,763	0.33	45	0.09	-141	40	0.41	167.9	10	0.12			
6,033	0.34	51	0.09	**-166**	**49**	**0.52**	177.4	11	0.13			

Table 1: The summary of results on the prostate cancer dataset by the five methods. The results of shrunken centroids and empirical Bayes methods are copied from Table 1 of Efron (2009). For Lasso-Logistic, MDR and SODA, the selected set with lowest BIC score is highlighted in bold font. ΔBIC: For MDR method, the difference of BIC_G between two adjacent steps. CVE: prediction error estimated by 10-fold cross-validation. #P: number of selected predictors. #M / #I: number of selected main effect and interaction terms by SODA.

$\boldsymbol{\theta}_{\mathcal{S}}$,

$$\mu_{i,k}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)=\Pr\left(y_i=k\mid\mathbf{X},\boldsymbol{\theta}_{\mathcal{S}}\right)=\frac{\exp\left(\boldsymbol{\theta}_{k,\mathcal{S}}^T\mathbf{z}_i\right)}{1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{z}_i\right)},\quad k=1,\ldots,K,$$

The true probability of $y_i=k$ with true parameters is $\mu_{i,k}\left(\boldsymbol{\theta}_0\right)$. Let $\sigma_{i,k}^2\left(\boldsymbol{\theta}_{\mathcal{S}}\right)$ denote the variance of $\mathbb{I}\left\{y_i=k\right\}$ with parameters $\boldsymbol{\theta}_{\mathcal{S}}$,

$$\sigma_{i,k}^2\left(\boldsymbol{\theta}_{\mathcal{S}}\right) \quad = \quad \mu_{i,k}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\left(1-\mu_{i,k}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right)$$

Let $\boldsymbol{s}_n\left(\boldsymbol{\theta}_{\mathcal{S}}\right)$ denote the score vector, where $\boldsymbol{s}_n\left(\boldsymbol{\theta}_{\mathcal{S}}\right)=\left[\boldsymbol{s}_{n,1}^T\left(\boldsymbol{\theta}_{\mathcal{S}}\right),\ldots,\boldsymbol{s}_{n,K-1}^T\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right]^T$ and

$$\boldsymbol{s}_{n,k}\left(\boldsymbol{\theta}_{\mathcal{S}}\right) \quad = \quad \frac{\partial l_n\left(\boldsymbol{\theta}_{\mathcal{S}}\right)}{\partial\boldsymbol{\theta}_{k,\mathcal{S}}}=\sum_{i=1}^{n}\left[\mathbb{I}\left\{y_i=k\right\}-\mu_{i,k}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right]\mathbf{z}_i.$$

Let $\mathbf{H}_n\left(\boldsymbol{\theta}_{\mathcal{S}}\right)$ denote the negative Hessian matrix of $l_n\left(\boldsymbol{\theta}_{\mathcal{S}}\right)$, which consists of $(K-1)\times(K-1)$ blocks. The k_1th row, k_2th column block is

$$\begin{aligned}\mathbf{H}_{n,k_1,k_2}\left(\boldsymbol{\theta}_{\mathcal{S}}\right) &= -\frac{\partial^2 l_n\left(\boldsymbol{\theta}_{\mathcal{S}}\right)}{\partial\boldsymbol{\theta}_{k_1,\mathcal{S}}\partial\boldsymbol{\theta}_{k_2,\mathcal{S}}^T}\\ &= \sum_{i=1}^{n}\mu_{i,k_1}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\left[\mathbb{I}\left\{k_1=k_2\right\}-\mu_{i,k_2}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right]\mathbf{z}_i\mathbf{z}_i^T\end{aligned}$$

With the notation of Kronecker product, $\mathbf{H}_n(\boldsymbol{\theta}_{\mathcal{S}})$ is

$$\mathbf{H}_n(\boldsymbol{\theta}_{\mathcal{S}}) = \sum_{i=1}^{n} \mathbf{U}_i(\boldsymbol{\theta}_{\mathcal{S}}) \otimes \left(\mathbf{z}_i \mathbf{z}_i^T\right),$$

where $\mathbf{U}_i(\boldsymbol{\theta}_{\mathcal{S}})$ is a $(K-1)\times(K-1)$ matrix and is a function of $\boldsymbol{\theta}_{\mathcal{S}}$. $\mathbf{U}_{i,k,k}(\boldsymbol{\theta}_{\mathcal{S}}) = \sigma^2_{i,k}(\boldsymbol{\theta}_{\mathcal{S}})$ and $\mathbf{U}_{i,k_1,k_2}(\boldsymbol{\theta}_{\mathcal{S}}) = -\mu_{i,k_1}(\boldsymbol{\theta}_{\mathcal{S}})\,\mu_{i,k_2}(\boldsymbol{\theta}_{\mathcal{S}})$ for $k_1 \neq k_2$. Let $\boldsymbol{s}_{\mathcal{S}}(\cdot)$ and $\mathbf{H}_{\mathcal{S}}(\cdot)$ respectively denote the sub-vector and sub-matrix of $\boldsymbol{s}_n(\cdot)$ and $\mathbf{H}_n(\cdot)$ corresponding to parameters associated terms in $\mathcal{S}$.

Let $\mathbf{Q}(\boldsymbol{\theta}_{\mathcal{S}})$ denote the Fisher information matrix. It consists of $(K-1)\times(K-1)$ blocks. The k_1th row, k_2th column block is

$$\begin{aligned}
\mathbf{Q}_{k_1,k_2}(\boldsymbol{\theta}_{\mathcal{S}}) &= \mathbb{E}\left[-\frac{\partial^2 l_1(\boldsymbol{\theta}_{\mathcal{S}})}{\partial \boldsymbol{\theta}_{k_1,\mathcal{S}} \partial \boldsymbol{\theta}_{k_2,\mathcal{S}}^T}\right] \\
&= \mathbb{E}\left[\left[\frac{\exp\left(\boldsymbol{\theta}_{k_1,\mathcal{S}}^T \mathbf{Z}\right)}{1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T \mathbf{Z}\right)}\right]\left[\mathbb{I}\{k_1=k_2\} - \frac{\exp\left(\boldsymbol{\theta}_{k_2,\mathcal{S}}^T \mathbf{Z}\right)}{1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T \mathbf{Z}\right)}\right]\mathbf{Z}\,\mathbf{Z}^T\right],
\end{aligned}$$

where the expectation is taken over sampling distribution of $(\mathbf{Z}, Y)$ under true parameters $\boldsymbol{\theta}_0$. Let $\mathbf{Q}_{\mathcal{S}}(\cdot)$ denote the sub-matrix of $\mathbf{Q}(\cdot)$ corresponding to parameters for terms in $\mathcal{S}$.

2.1 Proof of Lemmas 1~3

Lemmas 1~3 are required to prove the theorems in this article. Foygel and Drton (2011) proved similar lemmas for univariate generalized linear models with linear terms. Using similar arguments, we show the lemmas for the multinomial (multi-class) logistic regression model.

Lemma 1. *Under conditions C1 ~ C4, any local change in the Hessian is asymptotically bounded from above. Fix any positive integer constant Q and all $\mathcal{S}$ with $|\mathcal{S}| \leq Q$, for all $\boldsymbol{\theta}_{\mathcal{S}}$, $\boldsymbol{\theta}'_{\mathcal{S}}$, there exist constants $\lambda_3 > 0$ and $C_1 > 0$, such that for any constant integer $M > 2\kappa$, as $n \to \infty$,*

$$Pr\left\{\lambda_{\max}\left(\frac{1}{n}\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}}) - \frac{1}{n}\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}'_{\mathcal{S}})\right) \leq \lambda_3 \left\|\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{\theta}'_{\mathcal{S}}\right\|_2\right\} \geq 1 - C_1 n^{2\kappa - M} \to 1, \qquad (2)$$

where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a matrix.

Proof: Define $a_{i,k_1,k_2}(\boldsymbol{\theta}_{\mathcal{S}}) = \mu_{i,k_1}(\boldsymbol{\theta}_{\mathcal{S}})[\mathbb{I}\{k_1 = k_2\} - \mu_{i,k_2}(\boldsymbol{\theta}_{\mathcal{S}})]$. We have

$$\begin{aligned}
\left\|\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}}) - \mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}'_{\mathcal{S}})\right\|_{\text{sp}} &\leq \left\|\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}}) - \mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}'_{\mathcal{S}})\right\|_{\text{F}} \\
&= \sqrt{\sum_{k_1=1}^{K-1}\sum_{k_2=1}^{K-1}\left\|\mathbf{H}_{k_1,k_2,\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}}) - \mathbf{H}_{k_1,k_2,\mathcal{S}}(\boldsymbol{\theta}'_{\mathcal{S}})\right\|_{\text{F}}^2} \\
&\leq \sqrt{(K-1)^2 \max_{k_1,k_2}\left\|\mathbf{H}_{k_1,k_2,\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}}) - \mathbf{H}_{k_1,k_2,\mathcal{S}}(\boldsymbol{\theta}'_{\mathcal{S}})\right\|_{\text{F}}^2} \\
&= (K-1)\max_{k_1,k_2}\left\|\mathbf{H}_{k_1,k_2,\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}}) - \mathbf{H}_{k_1,k_2,\mathcal{S}}(\boldsymbol{\theta}'_{\mathcal{S}})\right\|_{\text{F}} \\
&= (K-1)\max_{k_1,k_2}\left\|\sum_{i=1}^{n} a_{i,k_1,k_2}(\boldsymbol{\theta}_{\mathcal{S}})\mathbf{z}_{i,\mathcal{S}}\mathbf{z}_{i,\mathcal{S}}^T - \sum_{i=1}^{n} a_{i,k_1,k_2}(\boldsymbol{\theta}'_{\mathcal{S}})\mathbf{z}_{i,\mathcal{S}}\mathbf{z}_{i,\mathcal{S}}^T\right\|_{\text{F}} \\
&= (K-1)\max_{k_1,k_2}\left\|\sum_{i=1}^{n}\left[a_{i,k_1,k_2}(\boldsymbol{\theta}_{\mathcal{S}}) - a_{i,k_1,k_2}(\boldsymbol{\theta}'_{\mathcal{S}})\right]\mathbf{z}_{i,\mathcal{S}}\mathbf{z}_{i,\mathcal{S}}^T\right\|_{\text{F}} \qquad (3)
\end{aligned}$$

Let $\boldsymbol{\theta}''_{\mathcal{S}} = t\boldsymbol{\theta}_{\mathcal{S}} + (1-t)\boldsymbol{\theta}'_{\mathcal{S}}$, then there exists $t \in (0,1)$ such that

$$\begin{aligned}
(3) &= (K-1)\max_{k_1,k_2}\left\|\sum_{i=1}^{n}\left\{\left[\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{\theta}'_{\mathcal{S}}\right]^T \nabla a_{i,k_1,k_2}(\boldsymbol{\theta}''_{\mathcal{S}})\right\}\mathbf{z}_{i,\mathcal{S}}\mathbf{z}_{i,\mathcal{S}}^T\right\|_{\text{F}} \\
&\leq (K-1)\max_{k_1,k_2}\left\|\sum_{i=1}^{n}\left\|\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{\theta}'_{\mathcal{S}}\right\|_2 \cdot \left\|\nabla a_{i,k_1,k_2}(\boldsymbol{\theta}''_{\mathcal{S}})\right\|_2 \cdot \mathbf{z}_{i,\mathcal{S}}\mathbf{z}_{i,\mathcal{S}}^T\right\|_{\text{F}} \\
&\leq \left\|\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{\theta}'_{\mathcal{S}}\right\|_2 \cdot (K-1)\max_{k_1,k_2}\sum_{i=1}^{n}\left\|\nabla a_{i,k_1,k_2}(\boldsymbol{\theta}''_{\mathcal{S}})\right\|_2 \cdot \left\|\mathbf{z}_{i,\mathcal{S}}\mathbf{z}_{i,\mathcal{S}}^T\right\|_{\text{F}} \qquad (4)
\end{aligned}$$

For $k_1 \neq k_2$, $a_{i,k_1,k_2}(\boldsymbol{\theta}_{\mathcal{S}}) = -\mu_{i,k_1}(\boldsymbol{\theta}_{\mathcal{S}})\mu_{i,k_2}(\boldsymbol{\theta}_{\mathcal{S}})$. Let $|\cdot|$ denote the element-wise absolute value of a vector. For h such that $h \neq k_1$ and $h \neq k_2$, and , then

$$\begin{aligned}
\left|\left[\nabla a_{i,k_1,k_2}(\boldsymbol{\theta}_{\mathcal{S}})\right]_h\right| &= \left|\frac{\partial a_{i,k_1,k_2}(\boldsymbol{\theta}_{\mathcal{S}})}{\partial \boldsymbol{\theta}_{h,\mathcal{S}}}\right| \\
&= \left|\frac{\exp\left(\boldsymbol{\theta}_{k_1,\mathcal{S}}^T\mathbf{z}_i\right)\exp\left(\boldsymbol{\theta}_{k_2,\mathcal{S}}^T\mathbf{z}_i\right)/\left[1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{z}_i\right)\right]^2}{\partial \boldsymbol{\theta}_{h,\mathcal{S}}}\right| \\
&= \left|\frac{2\left[1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{z}_i\right)\right]\exp\left(\boldsymbol{\theta}_{h,\mathcal{S}}^T\mathbf{z}_i\right)\exp\left(\boldsymbol{\theta}_{k_1,\mathcal{S}}^T\mathbf{z}_i\right)\exp\left(\boldsymbol{\theta}_{k_2,\mathcal{S}}^T\mathbf{z}_i\right)}{\left[1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{z}_i\right)\right]^4}\mathbf{z}_{i,\mathcal{S}}\right| \\
&\leq 2\left|\mathbf{z}_{i,\mathcal{S}}\right|,
\end{aligned}$$

where the last $\leq$ denotes element-wise $\leq$ of two vectors. Similarly we can show that for $h = k_1 \neq k_2$ or $h \neq k_1 = k_2$ or $h = k_1 = k_2$, we all have

$$\begin{aligned}
\left|\left[\nabla a_{i,k_1,k_2}(\boldsymbol{\theta}_{\mathcal{S}})\right]_h\right| &= \left|\frac{\partial a_{i,k_1,k_2}(\boldsymbol{\theta}_{\mathcal{S}})}{\partial \boldsymbol{\theta}_{h,\mathcal{S}}}\right| \\
&\leq 2\left|\mathbf{z}_{i,\mathcal{S}}\right|
\end{aligned}$$

Therefore $\left\|\nabla a_{i,k_1,k_2}\left(\boldsymbol{\theta}''_{\mathcal{S}}\right)\right\|_2 \leq 2\left(K-1\right)^{1/2}\left\|\mathbf{z}_{i,\mathcal{S}}\right\|_2$, and

$$\begin{aligned}
(4) &\leq 2\left(K-1\right)^{3/2}\cdot\left\|\boldsymbol{\theta}_{\mathcal{S}}-\boldsymbol{\theta}''_{\mathcal{S}}\right\|_2\cdot\sum_{i=1}^{n}\left\|\mathbf{z}_{i,\mathcal{S}}\right\|_2\cdot\left\|\mathbf{z}_{i,\mathcal{S}}\mathbf{z}_{i,\mathcal{S}}^T\right\|_{\mathrm{F}} \\
&\leq 2\left(K-1\right)^{3/2}\cdot\left\|\boldsymbol{\theta}_{\mathcal{S}}-\boldsymbol{\theta}'_{\mathcal{S}}\right\|_2\cdot\sum_{i=1}^{n}\left\|\mathbf{z}_{i,\mathcal{S}}\right\|_2^3 \\
&= 2\left(K-1\right)^{3/2}\cdot\left\|\boldsymbol{\theta}_{\mathcal{S}}-\boldsymbol{\theta}'_{\mathcal{S}}\right\|_2\cdot\sum_{i=1}^{n}\left[\sum_{j\in\mathcal{S}}z_{i,j}^2\right]^{3/2}. \qquad (5)
\end{aligned}$$

By Jensen's inequality,

$$\begin{aligned}
\sum_{i=1}^{n}\left[\sum_{j\in\mathcal{S}}z_{i,j}^2\right]^{3/2} &= |S|^{3/2}\sum_{i=1}^{n}\left[\frac{1}{|S|}\sum_{j\in\mathcal{S}}z_{i,j}^2\right]^{3/2} \\
&\leq |S|^{3/2}\sum_{i=1}^{n}\left[\frac{1}{|S|}\sum_{j\in\mathcal{S}}\left|z_{i,j}\right|^3\right] \\
&\leq |S|^{3/2}\max_{j\in\{1,\ldots,p\}}\sum_{i=1}^{n}\left|z_{i,j}\right|^3.
\end{aligned}$$

By condition C3, Z_j is sub-exponential for all j, so for any finite positive integer M, there exists a constant C_M such that

$$\mathbb{E}\left[\left|Z_j\right|^{6M}\right]\leq C_M, \quad \text{for all } j=1,\ldots,p. \qquad (6)$$

By Rosenthal's inequality, there is a constant R_M such that for all $j=1,\ldots,p$,

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{i=1}^{n}\left(\left|z_{i,j}\right|^3-\mathbb{E}\left[\left|Z_j\right|^3\right]\right)\right|^{2M}\right] \\
\leq\ & R_M\left\{\sum_{i=1}^{n}\mathbb{E}\left[\left(\left|z_{i,j}\right|^3-\mathbb{E}\left[\left|Z_j\right|^3\right]\right)^{2M}\right]+\left(\sum_{i=1}^{n}\mathbb{E}\left[\left(\left|z_{i,j}\right|^3-\mathbb{E}\left[\left|Z_j\right|^3\right]\right)^{2}\right]\right)^{M}\right\} \\
\leq\ & R_M\left[2^{2M}n\left\{\mathbb{E}\left[\left(\left|z_{i,j}\right|^3\right)^{2M}\right]+\left[\mathbb{E}\left(\left|Z_j\right|^3\right)\right]^{2M}\right\}+\left[4n\left\{\mathbb{E}\left[\left(\left|z_{i,j}\right|^6\right)\right]+\left[\mathbb{E}\left(\left|Z_j\right|^3\right)\right]^2\right\}\right]^{M}\right] \\
\leq\ & R_M\left[2^{2M}n\left\{C_{2M}+\left[\mathbb{E}\left(\left|Z_j\right|^3\right)\right]^{2M}\right\}+2^{2M}n^M\left[\left\{\mathbb{E}\left[\left(\left|z_{i,j}\right|^6\right)\right]+\left[\mathbb{E}\left(\left|Z_j\right|^3\right)\right]^2\right\}\right]^{M}\right] \\
\leq\ & C'_M n^M,
\end{aligned}$$

for sufficiently large n, where the positive constant C'_M is defined as

$$C'_M=1+2^{2M}\cdot R_M\cdot\max_j\left\{\left[\left\{\mathbb{E}\left[\left(\left|z_{i,j}\right|^6\right)\right]+\left[\mathbb{E}\left(\left|Z_j\right|^3\right)\right]^2\right\}\right]^{M}\right\}. \qquad (7)$$

For any positive constant integer M defined in 6, by Jensen's inequality,

$$\mathbb{E}\left(|Z_j|^3\right) \leq \left[\mathbb{E}\left(|Z_j|^{6M}\right)\right]^{1/2M} \leq C_M^{1/2M}. \tag{8}$$

therefore let constant $C_1 = C'_M/C_M$, then

$$\begin{aligned}
\Pr\left\{\sum_{i=1}^{n}|z_{i,j}|^3 > 2nC_M^{1/2M}\right\} &= \Pr\left\{\sum_{i=1}^{n}\left[|z_{i,j}|^3 - \mathbb{E}\left(|Z_j|^3\right)\right] > 2nC_M^{1/2M} - \sum_{i=1}^{n}\mathbb{E}\left(|Z_j|^3\right)\right\} \\
&\leq \Pr\left\{\sum_{i=1}^{n}\left[|z_{i,j}|^3 - \mathbb{E}\left(|Z_j|^3\right)\right] > 2nC_M^{1/2M} - nC_M^{1/2M}\right\} \\
&= \Pr\left\{\sum_{i=1}^{n}\left[|z_{i,j}|^3 - \mathbb{E}\left(|Z_j|^3\right)\right] > nC_M^{1/2M}\right\} \\
&= \Pr\left\{\left\{\sum_{i=1}^{n}\left[|z_{i,j}|^3 - \mathbb{E}\left(|Z_j|^3\right)\right]\right\}^{2M} > n^{2M}C_M\right\} \\
&\leq \frac{\mathbb{E}\left[\left\{\sum_{i=1}^{n}\left[|z_{i,j}|^3 - \mathbb{E}\left(|Z_j|^3\right)\right]\right\}^{2M}\right]}{n^{2M}C_M} \\
&\leq \frac{C'_M n^M}{C_M n^{2M}} \\
&= C_1 n^{-M}.
\end{aligned}$$

Let $q = \frac{1}{2}p\,(p+3)$ be the total number of terms in $\mathbf{Z}$, then with probability at least $1 - q \cdot C_1 n^{-M}$, there is a uniform bound,

$$\max_j \sum_{i=1}^{n} z_{i,j}^3 \leq 2nC_M^{1/2M},$$

and

$$\begin{aligned}
(5) &\leq 2\,(K-1)^{3/2}\left\|\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{\theta}'_{\mathcal{S}}\right\|_2 \cdot |S|^{3/2} \cdot \max_j \sum_{i=1}^{n} z_{i,j}^3. \\
&\leq 2\,(K-1)^{3/2}\left\|\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{\theta}'_{\mathcal{S}}\right\|_2 \cdot |S|^{3/2} \cdot 2nC_M^{1/2M} \\
&\leq \lambda_3 n\left\|\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{\theta}'_{\mathcal{S}}\right\|_2,
\end{aligned}$$

where positive constant $\lambda_3 = 4\,(K-1)^{3/2}\,Q^{3/2}C_M^{1/2M}$. By condition (C1), $q \leq n^{2\kappa}$, then

$$q \cdot C_1 n^{-M} \leq C_1 n^{2\kappa - M}.$$

Choose M as any constant integer with $M > 2\kappa$, then as $n \to \infty$, with probability at least $1 -$

$C_1 n^{2\kappa - M} \to 1$, uniformly for all $\mathcal{S}$ with $|\mathcal{S}| \leq Q$, for any $\boldsymbol{\theta}_{\mathcal{S}}$ and $\boldsymbol{\theta}'_{\mathcal{S}}$, there exists constant $\lambda_3 > 0$,

$$\lambda_{\max}\left(\frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right) - \frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}'_{\mathcal{S}}\right)\right) \leq \lambda_3 \left\|\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{\theta}'_{\mathcal{S}}\right\|_2.$$

Lemma 2. *Under conditions C1 ~ C4, eigenvalues of the Hessian are asymptotically bounded from above and below: Fix any positive constant integer Q. Choose positive constants $M > 2\kappa$ and $m > 2\kappa Q$. Then for any constant R such that $\forall \left\|\boldsymbol{\theta}_{\mathcal{S}}\right\|_2 \leq R$, there exist constants $\lambda_2 > \lambda_1 > 0$ and $C_1, C_2, C_3 > 0$, such that with probability at least $1 - C_1 n^{2\kappa - M} - C_2 n^{2\kappa - m} - C_3 n^{2\kappa Q - m} \to 1$ and uniformly for all $\mathcal{S}$ with $|\mathcal{S}| \leq Q$,*

$$\lambda_1 \leq \lambda_{\min}\left(\frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right) < \lambda_{\max}\left(\frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right) \leq \lambda_2.$$

Proof for lower bound: Let $\mathbf{u}$ denote a $|\mathcal{S}|$-length unit vector,

$$\begin{aligned}
\lambda_{\min}\left(\mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right) &= \min_{\|\mathbf{u}\|_2 = 1}\left\{\mathbf{u}^T \mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\mathbf{u}\right\} \\
&= \min_{\|\mathbf{u}\|_2 = 1}\left\{\mathbf{u}^T \frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\mathbf{u} + \mathbf{u}^T\left(\mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right) - \frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right)\mathbf{u}\right\} \\
&\leq \mathbf{v}^T \frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\mathbf{v} + \mathbf{v}^T\left(\mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right) - \frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right)\mathbf{v},
\end{aligned}$$

where $\mathbf{v}$ is an eigenvector of $\frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)$ corresponding to its lowest eigenvalue. Therefore,

$$\begin{aligned}
\lambda_{\min}\left(\frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right) &\geq \lambda_{\min}\left(\mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right) - \mathbf{v}^T\left(\mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right) - \frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right)\mathbf{v} \\
&\geq \lambda_{\min}\left(\mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right) - \lambda_{\max}\left(\mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right) - \frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right).
\end{aligned}$$

We first prove the positive-definiteness of $\mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)$, namely there exists constant $\lambda_1 > 0$ such that

$$\lambda_{\min}\left(\mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right) > \lambda_1.$$

By definition,

$$\mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right) = \mathbb{E}\left[\mathbf{U} \otimes \left(\mathbf{Z}_{\mathcal{S}}\mathbf{Z}_{\mathcal{S}}^T\right)\right] = \int \left[\mathbf{U} \otimes \left(\mathbf{Z}_{\mathcal{S}}\mathbf{Z}_{\mathcal{S}}^T\right)\right] dP\left(\mathbf{Z}_{\mathcal{S}}\right),$$

where $\mathbf{U}$ is a $(K-1) \times (K-1)$ matrix and is a function of $\boldsymbol{\theta}_{\mathcal{S}}$ and $\mathbf{Z}$,

$$\mathbf{U} = \boldsymbol{\Lambda}\left(\boldsymbol{\theta}_{\mathcal{S}}, \mathbf{Z}\right) - \left[\boldsymbol{\mu}\left(\boldsymbol{\theta}_{\mathcal{S}}, \mathbf{Z}\right)\right]\left[\boldsymbol{\mu}\left(\boldsymbol{\theta}_{\mathcal{S}}, \mathbf{Z}\right)\right]^T. \tag{9}$$

where $(K-1) \times (K-1)$ diagonal matrix $\boldsymbol{\Lambda}\left(\boldsymbol{\theta}_{\mathcal{S}}, \mathbf{Z}\right)$ has k-th diagonal element $\mu_k\left(\boldsymbol{\theta}_{\mathcal{S}}, \mathbf{Z}\right)$, and

vector $\boldsymbol{\mu}\left(\boldsymbol{\theta}_{\mathcal{S}}, \mathbf{Z}\right) = \left[\mu_1\left(\boldsymbol{\theta}_{\mathcal{S}}, \mathbf{Z}\right), \ldots, \mu_{K-1}\left(\boldsymbol{\theta}_{\mathcal{S}}, \mathbf{Z}\right)\right]^T$, where

$$\mu_k\left(\boldsymbol{\theta}_{\mathcal{S}}, \mathbf{Z}\right) = \frac{\exp\left(\boldsymbol{\theta}_{k,\mathcal{S}}^T \mathbf{Z}\right)}{1 + \sum_{l=1}^{K-1} \exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T \mathbf{Z}\right)}.$$

Let $\mathbf{v}$ denote the unit minimal eigenvector of $\mathrm{Cov}\left(\mathbf{Z}_{\mathcal{S}}\right)$. Without loss of generality, assume $\mathbb{E}\left[\mathbf{Z}\right] = \mathbf{0}$. By condition C4,

$$\mathbf{v}^T \mathbb{E}\left[\mathbf{Z}_{\mathcal{S}} \mathbf{Z}_{\mathcal{S}}^T\right] \mathbf{v} = \mathbf{v}^T \int \left[\mathbf{Z}_{\mathcal{S}} \mathbf{Z}_{\mathcal{S}}^T\right] dP\left(\mathbf{Z}_{\mathcal{S}}\right) \mathbf{v} \geq \tau_1.$$

Define subspace $\mathcal{R}_M = [-M, M]^{|\mathcal{S}|} \in \mathbb{R}^{|\mathcal{S}|}$ and complement $\mathcal{R}_M^c = \mathbb{R}^{|\mathcal{S}|} \backslash \mathcal{R}_M$, then

$$\begin{aligned}
\mathbf{v}^T \mathbb{E}\left[\mathbf{Z}_{\mathcal{S}} \mathbf{Z}_{\mathcal{S}}^T\right] \mathbf{v} &= \mathbf{v}^T \int_{\mathbf{Z}_{\mathcal{S}} \in \mathcal{R}_M} \left[\mathbf{Z}_{\mathcal{S}} \mathbf{Z}_{\mathcal{S}}^T\right] dP\left(\mathbf{Z}_{\mathcal{S}}\right) \mathbf{v} + \mathbf{v}^T \int_{\mathbf{Z}_{\mathcal{S}} \in \mathcal{R}_M^c} \left[\mathbf{Z}_{\mathcal{S}} \mathbf{Z}_{\mathcal{S}}^T\right] dP\left(\mathbf{Z}_{\mathcal{S}}\right) \mathbf{v} \\
&= f_1(M) + f_2(M) \\
&\geq f_1(M),
\end{aligned}$$

where functions $f_1(M)$ and $f_2(M)$ are defined as

$$f_1(M) = \mathbf{v}^T \int_{\mathbf{Z}_{\mathcal{S}} \in \mathcal{R}_M} \left[\mathbf{Z}_{\mathcal{S}} \mathbf{Z}_{\mathcal{S}}^T\right] dP\left(\mathbf{Z}_{\mathcal{S}}\right) \mathbf{v}, \tag{10}$$

$$f_2(M) = \mathbf{v}^T \int_{\mathbf{Z}_{\mathcal{S}} \in \mathcal{R}_M^c} \left[\mathbf{Z}_{\mathcal{S}} \mathbf{Z}_{\mathcal{S}}^T\right] dP\left(\mathbf{Z}_{\mathcal{S}}\right) \mathbf{v}. \tag{11}$$

Because of the semi-positive definiteness of $\left[\mathbf{Z}_{\mathcal{S}} \mathbf{Z}_{\mathcal{S}}^T\right]$, $f_1(M)$ is a increasing function and $f_2(M)$ is a decreasing function of M, and $f_1(0) = 0$ and $f_1(+\infty) = \tau_1$. There exists constant $M > 0$ such that

$$f_1(M) = \mathbf{v}^T \int_{\mathbf{Z}_{\mathcal{S}} \in \mathcal{R}_M} \left[\mathbf{Z}_{\mathcal{S}} \mathbf{Z}_{\mathcal{S}}^T\right] dP\left(\mathbf{Z}_{\mathcal{S}}\right) \mathbf{v} > \frac{1}{2}\tau_1.$$

For $\mathbf{Z} \in \mathcal{R}_M$, it is straightforward to show that there exists positive constant $C_M > 0$ such that $C_M < \sum_{k=1}^{K-1} \mu_k\left(\boldsymbol{\theta}_{\mathcal{S}}, \mathbf{Z}\right) < 1 - C_M$ and $C_M < \mu_k\left(\boldsymbol{\theta}_{\mathcal{S}}, \mathbf{Z}\right) < 1 - C_M$ for $k = 1, \ldots, K-1$.

Let $\mathbf{U}_{k_1,k_2}$ denote the k_1th row, k_2th column element of $\mathbf{U}$. By definition of $\mathbf{U}$ in (9),

$$\begin{aligned}
|\mathbf{U}_{k_1,k_1}| - \sum_{k_2 \neq k_1} |\mathbf{U}_{k_1,k_2}| &= \mu_{k_1}\left(\boldsymbol{\theta}_{\mathcal{S}}, \mathbf{Z}\right) - \mu_{k_1}\left(\boldsymbol{\theta}_{\mathcal{S}}, \mathbf{Z}\right) \sum_{k_2=1}^{K-1} \mu_{k_2}\left(\boldsymbol{\theta}_{\mathcal{S}}, \mathbf{Z}\right) && (12) \\
&= \mu_{k_1}\left(\boldsymbol{\theta}_{\mathcal{S}}, \mathbf{Z}\right) \left[1 - \sum_{k_2=1}^{K-1} \mu_{k_2}\left(\boldsymbol{\theta}_{\mathcal{S}}, \mathbf{Z}\right)\right] && (13) \\
&\geq C_M^2, && (14)
\end{aligned}$$

therefore $\mathbf{U} - \frac{1}{2} C_M^2 \mathbf{I}_{K-1} \succeq 0$ and $\mathbf{U} \succeq \frac{1}{2} C_M^2 \mathbf{I}_{K-1}$, since $\mathbf{U} - \frac{1}{2} C_M^2 \mathbf{I}_{K-1}$ is a diagonally dominant

matrix, where $\mathbf{I}_{K-1}$ is an identify matrix. By property of Kronecker product,

$$\mathbf{U} \otimes \left(\mathbf{Z}_{\mathcal{S}}\mathbf{Z}_{\mathcal{S}}^T\right) \succeq \frac{1}{2}C_M^2\mathbf{I}_{K-1} \otimes \left(\mathbf{Z}_{\mathcal{S}}\mathbf{Z}_{\mathcal{S}}^T\right). \tag{15}$$

Finally,

$$\begin{aligned}
\min_{\|\mathbf{u}\|=1} \mathbf{u}^T\mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\mathbf{u} &= \min_{\|\mathbf{u}\|=1} \mathbf{u}^T\left\{\int\left[\mathbf{U} \otimes \left(\mathbf{Z}_{\mathcal{S}}\mathbf{Z}_{\mathcal{S}}^T\right)\right] dP\left(\mathbf{Z}_{\mathcal{S}}\right)\right\}\mathbf{u} \\
&= \min_{\|\mathbf{u}\|=1} \int\left[\mathbf{u}^T\mathbf{U} \otimes \left(\mathbf{Z}_{\mathcal{S}}\mathbf{Z}_{\mathcal{S}}^T\right)\mathbf{u}\right] dP\left(\mathbf{Z}_{\mathcal{S}}\right) \\
&\geq \min_{\|\mathbf{u}\|=1} \int_{\mathbf{Z}_{\mathcal{S}}\in\mathcal{R}_M}\left[\mathbf{u}^T\mathbf{U} \otimes \left(\mathbf{Z}_{\mathcal{S}}\mathbf{Z}_{\mathcal{S}}^T\right)\mathbf{u}\right] dP\left(\mathbf{Z}_{\mathcal{S}}\right) \\
&\geq \min_{\|\mathbf{u}\|=1} \int_{\mathbf{Z}_{\mathcal{S}}\in\mathcal{R}_M}\left[\mathbf{u}^T\frac{1}{2}C_M^2\mathbf{I}_{K-1} \otimes \left(\mathbf{Z}_{\mathcal{S}}\mathbf{Z}_{\mathcal{S}}^T\right)\mathbf{u}\right] dP\left(\mathbf{Z}_{\mathcal{S}}\right) \\
&\geq \frac{1}{2}C_M^2 \min_{\|\mathbf{u}\|=1} \int_{\mathbf{Z}_{\mathcal{S}}\in\mathcal{R}_M}\left[\mathbf{u}^T\mathbf{I}_{K-1} \otimes \left(\mathbf{Z}_{\mathcal{S}}\mathbf{Z}_{\mathcal{S}}^T\right)\mathbf{u}\right] dP\left(\mathbf{Z}_{\mathcal{S}}\right) \\
&= \frac{1}{2}C_M^2 \min_{\|\mathbf{u}\|=1} \mathbf{u}^T\mathbf{I}_{K-1} \otimes \int_{\mathbf{Z}_{\mathcal{S}}\in\mathcal{R}_M}\left[\left(\mathbf{Z}_{\mathcal{S}}\mathbf{Z}_{\mathcal{S}}^T\right)\right] dP\left(\mathbf{Z}_{\mathcal{S}}\right)\mathbf{u} \\
&= \frac{1}{2}C_M^2 \cdot \frac{1}{2}\tau_1
\end{aligned}$$

thus there exists constant $\lambda_1 = C_M^2\tau_1/4 > 0$,

$$\lambda_{\min}\left(\mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right) > \lambda_1.$$

Next we derive the bound on $\lambda_{\max}\left(\mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right) - \frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right)$. Note that

$$\begin{aligned}
\lambda_{\max}^2\left(\mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right) - \frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right) &= \left\|\mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right) - \frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right\|_{\text{sp}}^2 \\
&\leq \left\|\mathbf{Q}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right) - \frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right\|_{\text{F}}^2 \\
&= \sum_{k_1=1}^{K-1}\sum_{k_2=1}^{K-1}\sum_{j_1\in\mathcal{S}}\sum_{j_2\in\mathcal{S}} W_{k_1,k_2,j_1,j_2}^2,
\end{aligned}$$

where

$$\begin{aligned}
W_{k_1,k_2,j_1,j_2} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\exp\left(\boldsymbol{\theta}_{k_1,\mathcal{S}}^T\mathbf{z}_i\right)}{1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{z}_i\right)}\right]\left[\mathbb{I}\left\{k_1=k_2\right\} - \frac{\exp\left(\boldsymbol{\theta}_{k_2,\mathcal{S}}^T\mathbf{z}_i\right)}{1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{z}_i\right)}\right] z_{i,j_1}z_{i,j_2} \\
&\quad -\mathbb{E}\left\{\left[\frac{\exp\left(\boldsymbol{\theta}_{k_1,\mathcal{S}}^T\mathbf{Z}\right)}{1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{Z}\right)}\right]\left[\mathbb{I}\left\{k_1=k_2\right\} - \frac{\exp\left(\boldsymbol{\theta}_{k_2,\mathcal{S}}^T\mathbf{Z}\right)}{1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{Z}\right)}\right] Z_{j_1}Z_{j_2}\right\}.
\end{aligned}$$

Following the technique used in inequality (81) of Ravikumar et al. (2011), we derive a moment

bound for W_{k_1,k_2,j_1,j_2}. Define W_{i,k_1,k_2,j_1,j_2} with

$$\begin{aligned} W_{i,k_1,k_2,j_1,j_2} &= \left[\frac{\exp\left(\boldsymbol{\theta}_{k_1,\mathcal{S}}^T\mathbf{z}_i\right)}{1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{z}_i\right)}\right]\left[\mathbb{I}\{k_1=k_2\}-\frac{\exp\left(\boldsymbol{\theta}_{k_2,\mathcal{S}}^T\mathbf{z}_i\right)}{1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{z}_i\right)}\right]z_{i,j_1}z_{i,j_2} \\ &\quad -\mathbb{E}\left\{\left[\frac{\exp\left(\boldsymbol{\theta}_{k_1,\mathcal{S}}^T\mathbf{Z}\right)}{1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{Z}\right)}\right]\left[\mathbb{I}\{k_1=k_2\}-\frac{\exp\left(\boldsymbol{\theta}_{k_2,\mathcal{S}}^T\mathbf{Z}\right)}{1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{Z}\right)}\right]Z_{j_1}Z_{j_2}\right\}. \end{aligned}$$

Thus $W_{k_1,k_2,j_1,j_2}=\sum_{i=1}^n W_{i,k_1,k_2,j_1,j_2}/n$. Let m be a positive integer, by Rosenthal's inequality, there exists a constant C_m such that

$$\mathbb{E}\left[\left(\sum_{i=1}^n W_{i,k_1,k_2,j_1,j_2}\right)^{2m}\right] \leq C_m\left[\sum_{i=1}^n\mathbb{E}\left[\left(W_{i,k_1,k_2,j_1,j_2}\right)^{2m}\right]+\left\{\sum_{i=1}^n\mathbb{E}\left[\left(W_{i,k_1,k_2,j_1,j_2}\right)^2\right]\right\}^m\right].$$

For the first set of terms, we know

$$\begin{aligned} &\mathbb{E}\left[\left(W_{i,k_1,k_2,j_1,j_2}\right)^{2m}\right] \\ \leq\ & 2^{2m}\left\{\mathbb{E}\left[\left(\left[\frac{\exp\left(\boldsymbol{\theta}_{k_1,\mathcal{S}}^T\mathbf{z}_i\right)}{1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{z}_i\right)}\right]\left[\mathbb{I}\{k_1=k_2\}-\frac{\exp\left(\boldsymbol{\theta}_{k_2,\mathcal{S}}^T\mathbf{z}_i\right)}{1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{z}_i\right)}\right]z_{i,j_1}z_{i,j_2}\right)^{2m}\right]\right\} \\ &+2^{2m}\mathbb{E}^{2m}\left\{\left[\frac{\exp\left(\boldsymbol{\theta}_{k_1,\mathcal{S}}^T\mathbf{Z}\right)}{1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{Z}\right)}\right]\left[\mathbb{I}\{k_1=k_2\}-\frac{\exp\left(\boldsymbol{\theta}_{k_2,\mathcal{S}}^T\mathbf{Z}\right)}{1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{Z}\right)}\right]Z_{j_1}Z_{j_2}\right\} \\ \leq\ & 2^{2m}\left\{\mathbb{E}\left[\left(z_{i,j_1}z_{i,j_2}\right)^{2m}\right]+\left[\mathbb{E}\left(Z_{j_1}Z_{j_2}\right)\right]^{2m}\right\} \\ \leq\ & 2^{2m}\left\{\sqrt{\mathbb{E}\left[\left(z_{i,j_1}\right)^{4m}\right]\mathbb{E}\left[\left(z_{i,j_2}\right)^{4m}\right]}+\left[\mathbb{E}\left(Z_{j_1}^2\right)\mathbb{E}\left(Z_{j_2}^2\right)\right]^m\right\} \\ \leq\ & C_{m,1}, \end{aligned}$$

by sub-exponential tail condition for each Z_j, $j=1,\ldots,p$, where $C_{m,1}$ is a positive constant depending only on m.

For the second set of terms, by taking $m=1$, we can also show that there exists a constant $C_{m,2}$ depending only on m,

$$\mathbb{E}\left[\left(W_{i,k_1,k_2,j_1,j_2}\right)^2\right]\leq C_{m,2}.$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[\left(W_{k_1,k_2,j_1,j_2}\right)^{2m}\right] &= \frac{1}{n^{2m}}\mathbb{E}\left[\left(\sum_{i=1}^n W_{i,k_1,k_2,j_1,j_2}\right)^{2m}\right] \\ &\leq \frac{1}{n^{2m}}C_m\left(nC_{m,1}+n^mC_{m,2}^m\right) \\ &\leq C_m'n^{-m}, \end{aligned}$$

where C'_m is a constant depending only on m.

Let $\{m_{k_1,k_2,j_1,j_2} : k_1, k_2 = 1, \ldots, K-1,\ j_1, j_2 \in \mathcal{S}\}$ be a set non-negative integers with

$$\sum_{k_1=1}^{K-1}\sum_{k_2=1}^{K-1}\sum_{j_1\in\mathcal{S}}\sum_{j_2\in\mathcal{S}} m_{k_1,k_2,j_1,j_2} = m, \tag{16}$$

then by iteratively applying Cauchy-Schwarz inequality, we can show that there exists a constant C_2 such that

$$\mathbb{E}\left[\prod_{k_1=1}^{K-1}\prod_{k_2=1}^{K-1}\prod_{j_1\in\mathcal{S}}\prod_{j_2\in\mathcal{S}} W_{k_1,k_2,j_1,j_2}^{2m_{k_1,k_2,j_1,j_2}}\right] \;\leq\; C_2 n^{-m}. \tag{17}$$

For example, suppose there are three W variables, $W_1^{2m_1}$, $W_2^{2m_2}$ and $W_3^{2m_3}$ with $m_1+m_2+m_3 = m$, then

$$\begin{aligned}
\mathbb{E}\left[W_1^{2m_1}\cdot W_2^{2m_2}\cdot W_3^{2m_3}\right] &\leq \sqrt{\mathbb{E}\left[W_1^{4m_1}\right]\mathbb{E}\left[\left(W_2^{4m_2}\cdot W_3^{4m_3}\right)\right]} \\
&\leq \sqrt{\mathbb{E}\left[W_1^{4m_1}\right]\sqrt{\mathbb{E}\left[\left(W_2^{8m_2}\right)\right]\mathbb{E}\left[\left(W_3^{8m_3}\right)\right]}} \\
&\leq \sqrt{C_{m_1}n^{-2m_1}\sqrt{C_{m_2}n^{-4m_2}\cdot C_{m_3}n^{-4m_3}}} \\
&= \sqrt{C_{m_1}\sqrt{C_{m_2}C_{m_3}}}n^{-m_1}n^{-m_2}n^{-m_3} \\
&= \sqrt{C_{m_1}\sqrt{C_{m_2}C_{m_3}}}n^{-m},
\end{aligned}$$

where $\sqrt{C_{m_1}\sqrt{C_{m_2}C_{m_3}}}$ is a constant depending only on m_1, m_2 and m_3.

Then

$$\begin{aligned}
\Pr\left\{\lambda_{\max}^2\left(\mathbf{Q}_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}}) - \frac{1}{n}\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}})\right) > \frac{1}{2}\lambda_1\right\} &\leq \Pr\left\{\left\|\mathbf{Q}_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}}) - \frac{1}{n}\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}})\right\|_{\mathrm{F}}^2 > \frac{1}{2}\lambda_1\right\} \\
&= \Pr\left\{\left\|\mathbf{Q}_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}}) - \frac{1}{n}\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}})\right\|_{\mathrm{F}}^{2m} > \frac{1}{2^m}\lambda_1^m\right\} \\
&\leq \frac{\mathbb{E}\left\{\left\|\mathbf{Q}_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}}) - \frac{1}{n}\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}})\right\|_{\mathrm{F}}^{2m}\right\}}{\lambda_1^m/2^m} \\
&= \left(\frac{2}{\lambda_1}\right)^m \mathbb{E}\left\{\sum_{k_1=1}^{K-1}\sum_{k_2=1}^{K-1}\sum_{j_1\in\mathcal{S}}\sum_{j_2\in\mathcal{S}} W_{k_1,k_2,j_1,j_2}^2\right\}^m \\
&\leq \left(\frac{2}{\lambda_1}\right)^m C_2\binom{m+(K-1)^2Q^2-1}{(K-1)^2Q^2-1}n^{-m},
\end{aligned}$$

since there are $\binom{m+(K-1)^2Q^2-1}{(K-1)^2Q^2-1}$ expectations in the summation after expanding the $\{\cdot\}^m$ and each of them is $\leq C_2 n^{-m}$ by (17). Define $C_3 = \left(\frac{2}{\lambda_1}\right)^m C_2 \binom{m+(K-1)^2Q^2-1}{(K-1)^2Q^2-1}$, then with probability at least $1 - C_3 n^{-m}$,

$$\lambda_{\max}^2 \left(\mathbf{Q}_{\mathcal{S}} \left(\boldsymbol{\theta}_{\mathcal{S}} \right) - \frac{1}{n} \mathbf{H}_{\mathcal{S}} \left(\boldsymbol{\theta}_{\mathcal{S}} \right) \right) \leq \frac{1}{2} \lambda_1,$$

and

$$\begin{aligned} \lambda_{\min} \left(\frac{1}{n} \mathbf{H}_{\mathcal{S}} \left(\boldsymbol{\theta}_{\mathcal{S}} \right) \right) &\geq \lambda_{\min} \left(\mathbf{Q}_{\mathcal{S}} \left(\boldsymbol{\theta}_{\mathcal{S}} \right) \right) - \lambda_{\max} \left(\mathbf{Q}_{\mathcal{S}} \left(\boldsymbol{\theta}_{\mathcal{S}} \right) - \frac{1}{n} \mathbf{H}_{\mathcal{S}} \left(\boldsymbol{\theta}_{\mathcal{S}} \right) \right) \\ &\geq \lambda_1 - \frac{1}{2} \lambda_1 \\ &= \frac{1}{2} \lambda_1 > 0. \end{aligned}$$

Let $q = \frac{1}{2} p \left(p + 3 \right)$, then with probability at least $1 - C_3 q^Q n^{-m}$, uniformly for all $\mathcal{S}$ with $|\mathcal{S}| \leq Q$, $\lambda_{\min} \left(\frac{1}{n} \mathbf{H}_{\mathcal{S}} \left(\boldsymbol{\theta}_{\mathcal{S}} \right) \right) \geq \frac{1}{2} \lambda_1 > 0$. Choose $m > 2\kappa Q$, then as $n \to \infty$,

$$1 - C_3 q^Q n^{-m} \geq 1 - C_3 n^{2\kappa Q - m} \to 1,$$

that proves the lower bound.

Proof for upper bound: By Lemma 1, for all $\mathcal{S}$ with $|\mathcal{S}| \leq Q$ and $\boldsymbol{\theta}_{\mathcal{S}}$, there exists a constant $C_1 > 0$, as $n \to \infty$, with probability at least $1 - C_1 n^{2\kappa - M} \to 1$,

$$\begin{aligned} \lambda_{\max} \left(\frac{1}{n} \mathbf{H}_{\mathcal{S}} \left(\boldsymbol{\theta}_{\mathcal{S}} \right) \right) &\leq \lambda_{\max} \left(\frac{1}{n} \mathbf{H}_{\mathcal{S}} \left(\boldsymbol{\theta}_{\mathcal{S}} \right) - \frac{1}{n} \mathbf{H}_{\mathcal{S}} \left(\mathbf{0} \right) \right) + \lambda_{\max} \left(\frac{1}{n} \mathbf{H}_{\mathcal{S}} \left(\mathbf{0} \right) \right) \\ &\leq \lambda_3 \left\| \boldsymbol{\theta}_{\mathcal{S}} \right\|_2 + \lambda_{\max} \left(\frac{1}{n} \mathbf{H}_{\mathcal{S}} \left(\mathbf{0} \right) \right). \end{aligned}$$

We derive a bound for $\lambda_{\max} \left(\frac{1}{n} \mathbf{H}_{\mathcal{S}} \left(\mathbf{0} \right) \right)$. For any $\mathcal{S}$ with $|\mathcal{S}| \leq Q$,

$$\begin{aligned} \left\| \mathbf{H}_{\mathcal{S}} \left(\mathbf{0} \right) \right\|_{\text{sp}} &\leq \left\| \mathbf{H}_{\mathcal{S}} \left(\mathbf{0} \right) \right\|_{\text{F}} \\ &= (K-1) \max_{k_1, k_2} \left\| \mathbf{H}_{k_1, k_2, \mathcal{S}} \left(\mathbf{0} \right) \right\|_{\text{F}} \\ &= (K-1) \max_{k_1, k_2} \left\| \sum_{i=1}^{n} a_{i, k_1, k_2} \left(\mathbf{0} \right) \mathbf{z}_{i, \mathcal{S}} \mathbf{z}_{i, \mathcal{S}}^T \right\|_{\text{F}} \\ &= (K-1) \max_{k_1, k_2} \left\| \sum_{i=1}^{n} \left[\mu_{i, k_1} \left(\boldsymbol{\theta}_{\mathcal{S}} \right) \left[\mathbb{I} \left\{ k_1 = k_2 \right\} - \mu_{i, k_2} \left(\boldsymbol{\theta}_{\mathcal{S}} \right) \right] \right] \mathbf{z}_{i, \mathcal{S}} \mathbf{z}_{i, \mathcal{S}}^T \right\|_{\text{F}} \\ &\leq (K-1) \left\| \sum_{i=1}^{n} \mathbf{z}_{i, \mathcal{S}} \mathbf{z}_{i, \mathcal{S}}^T \right\|_{\text{F}} \\ &\leq (K-1) \sum_{i=1}^{n} \sum_{j \in \mathcal{S}} z_{i,j}^2 \\ &\leq (K-1) Q \max_{j \in \{1, \ldots, p\}} \sum_{i=1}^{n} z_{i,j}^2. \end{aligned}$$

Using the similar technique of proof of Lemma 1, we derive a bound for $\max_{j\in\{1,\ldots,p\}}\sum_{i=1}^{n} z_{i,j}^2$. By Rosenthal's inequality,

$$\begin{aligned}\mathbb{E}\left[\left|\sum_{i=1}^{n}\left(z_{i,j}^2-\mathbb{E}\left[Z_j^2\right]\right)\right|^{2m}\right] &\leq R_m\left[\sum_{i=1}^{n}\mathbb{E}\left[\left(z_{i,j}^2-\mathbb{E}\left[Z_j^2\right]\right)^{2m}\right]+\left\{\sum_{i=1}^{n}\mathbb{E}\left[\left(z_{i,j}^2-\mathbb{E}\left[Z_j^2\right]\right)^{2}\right]\right\}^{m}\right]\\ &\leq C_2 n^m,\end{aligned}$$

where C_1 is a positive constant independent of n. Since Z_j is sub-exponential, for any finite positive integer m, there exists a constant C_m such that

$$\mathbb{E}\left(Z_j^{4m}\right)\leq C_m,\quad \text{for all } j=1,\ldots,p. \tag{18}$$

By Jensen's inequality,

$$\mathbb{E}\left(Z_j^2\right)\leq\left[\mathbb{E}\left(Z_j^{4m}\right)\right]^{1/2m}\leq C_m^{1/2m}. \tag{19}$$

Therefore for any $j\in\{1,\ldots,p\}$,

$$\begin{aligned}\Pr\left\{\sum_{i=1}^{n} z_{i,j}^2>2nC_m^{1/2m}\right\} &= \Pr\left\{\sum_{i=1}^{n}\left[z_{i,j}^2-\mathbb{E}\left(Z_j^2\right)\right]>2nC_m^{1/2m}-n\mathbb{E}\left[Z_j\right]\right\}\\ &\leq \Pr\left\{\sum_{i=1}^{n}\left[z_{i,j}^2-\mathbb{E}\left(Z_j^2\right)\right]>2nC_m^{1/2m}-nC_m^{1/2m}\right\}\\ &\leq \Pr\left\{\sum_{i=1}^{n}\left[z_{i,j}^2-\mathbb{E}\left(Z_j^2\right)\right]>nC_m^{1/2m}\right\}\\ &= \Pr\left\{\left\{\sum_{i=1}^{n}\left[z_{i,j}^2-\mathbb{E}\left(Z_j^2\right)\right]\right\}^{2m}>n^{2m}C_m\right\}\\ &\leq \frac{\mathbb{E}\left[\left\{\sum_{i=1}^{n}\left[z_{i,j}^2-\mathbb{E}\left(Z_j^2\right)\right]\right\}^{2m}\right]}{n^{2m}C_m}\\ &\leq \frac{C_1 n^m.}{C_m n^{2m}}\\ &= C_2 n^{-m},\end{aligned}$$

where $C_2=C_1/C_m$. Therefore with probability at least $1-qC_2n^{-m}$, there is a uniform bound,

$$\max_{j\in\{1,\ldots,p\}}\sum_{i=1}^{n} z_{i,j}^3\leq 2nC_m^{1/2m}, \tag{20}$$

and uniformly for all $\mathcal{S}$ with $|\mathcal{S}|\leq Q$,

$$\begin{aligned}\left\|\mathbf{H}_{\mathcal{S}}\left(\mathbf{0}\right)\right\|_{\text{sp}} &\leq (K-1)\,Q\max_{j\in\{1,\ldots,p\}}\sum_{i=1}^{n} z_{i,j}^2 &(21)\\ &\leq 2\,(K-1)\,QC_m^{1/2m}\cdot n. &(22)\end{aligned}$$

Finally we have for any $\|\boldsymbol{\theta}_{\mathcal{S}}\|_2 \leq R$,

$$\begin{aligned}\lambda_{\max}\left(\frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right) &\leq \lambda_3\left\|\boldsymbol{\theta}_{\mathcal{S}}\right\|_2 + \lambda_{\max}\left(\frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\mathbf{0}\right)\right)\\ &\leq \lambda_3 R + 2\left(K-1\right)QC_m^{1/2m}.\end{aligned}$$

Choose any integer $m > 2\kappa$, and define constant $\lambda_2 = \lambda_3 R + 2\left(K-1\right)QC_m^{1/2m}$, then as $n \to \infty$, with probability at least

$$1 - C_1 n^{2\kappa - M} - qC_2 n^{-m} \geq 1 - C_1 n^{2\kappa - M} - C_2 n^{2\kappa - m} \to 1, \tag{23}$$

for all $\mathcal{S}$ with $|\mathcal{S}| \leq Q$,

$$\lambda_{\max}\left(\frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right) \leq \lambda_2. \tag{24}$$

Merging the results for lower and upper bounds, there exist positive constants C_1, C_2 and C_3 such that as $n \to \infty$, choose $M > 4\kappa$ and $m > 2\kappa Q$, then with probability at least

$$1 - C_1 n^{2\kappa - M} - C_2 n^{2\kappa - m} - C_3 n^{2\kappa Q - m} \to 1, \tag{25}$$

we have

$$\lambda_1 \leq \lambda_{\min}\left(\frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right) < \lambda_{\max}\left(\frac{1}{n}\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}\right)\right) \leq \lambda_2. \tag{26}$$

Lemma 3 is related to Lemma 1(i) of Foygel and Drton (2011) for multinomial logistic regression ($K > 2$).

Lemma 3. *Fix any positive constant $Q > 0$. Under conditions C1 ~ C4, as $n \to \infty$, there exists constants $r_1, r_2 > 0$, with probability at least $1 - n^{-r_1}q^{-r_2}$, uniformly for all $\mathcal{S} \supsetneq \mathcal{A}$ with $|\mathcal{S}| \leq Q$,*

$$l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right) - l_n\left(\boldsymbol{\theta}_0\right) \leq \left[|\mathcal{S}| - |\mathcal{A}|\right]\log\left(n^{r_1}q^{1+r_2}\right) + \varepsilon_n,$$

where $\boldsymbol{\theta}_0$ is the true parameter vector, $q = p\left(p+3\right)/2$ and $\varepsilon_n = o\left(n^{-1/3}\right)$.

Proof: There exists $t \in (0,1)$, and $\boldsymbol{\theta}_{\mathcal{S}}^* = t\tilde{\boldsymbol{\theta}}_{\mathcal{S}} + \left(1-t\right)\boldsymbol{\theta}_0$, such that

$$\begin{aligned}& l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right) - l_n\left(\boldsymbol{\theta}_0\right)\\ =\ & \left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)^T \boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right) - \frac{1}{2}\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)^T \mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}^*\right)\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)\\ =\ & \left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)^T \boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right) - \frac{1}{2}\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)^T \mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right)\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)\\ & + \frac{1}{2}\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)^T \left[\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right) - \mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{\mathcal{S}}^*\right)\right]\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right).\end{aligned} \tag{27}$$

By Lemma 1, with probability at least $1 - C_1 n^{2\kappa - M}$,

$$\begin{aligned} (27) &\leq \left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)^T \boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right) - \frac{1}{2}\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)^T \mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right)\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right) + \frac{1}{2}\lambda_3 n \left\|\boldsymbol{\theta}^*_{\mathcal{S}} - \boldsymbol{\theta}_0\right\|_2^3 \\ &= \left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)^T \boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right) - \frac{1}{2}\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)^T \mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right)\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right) + \varepsilon_n. \end{aligned}$$

where $\varepsilon_n = \frac{1}{2}\lambda_3 n \left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right\|_2^3 = o\left(n^{-1/3}\right)$ by Theorem 1.

Maximizing

$$\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)^T \boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right) - \frac{1}{2}\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)^T \mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right)\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right) \tag{28}$$

with respect to $\tilde{\boldsymbol{\theta}}_{\mathcal{S}}$, we get

$$l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right) - l_n\left(\boldsymbol{\theta}_0\right) \leq \frac{1}{2}\boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right)^T \mathbf{H}_{\mathcal{S}}^{-1}\left(\boldsymbol{\theta}_0\right)\boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right) + \varepsilon_n.$$

Define $\tilde{\mathbf{y}}$ and $\tilde{\boldsymbol{\mu}}$ as concatenated vectors of length $n\left(K-1\right)$, such that

$$\tilde{\mathbf{y}} = \begin{bmatrix} \tilde{\mathbf{y}}_1 \\ \tilde{\mathbf{y}}_2 \\ \vdots \\ \tilde{\mathbf{y}}_{K-1} \end{bmatrix}, \quad \tilde{\mathbf{y}}_k = \begin{bmatrix} \mathbb{I}\left\{y_1 = k\right\} \\ \mathbb{I}\left\{y_2 = k\right\} \\ \vdots \\ \mathbb{I}\left\{y_n = k\right\} \end{bmatrix}, \quad 1 \leq k \leq K-1,$$

and

$$\tilde{\boldsymbol{\mu}} = \begin{bmatrix} \tilde{\boldsymbol{\mu}}_1 \\ \tilde{\boldsymbol{\mu}}_2 \\ \vdots \\ \tilde{\boldsymbol{\mu}}_{K-1} \end{bmatrix}, \quad \tilde{\boldsymbol{\mu}}_k = \begin{bmatrix} \mu_{1,k}\left(\boldsymbol{\theta}_0\right) \\ \mu_{2,k}\left(\boldsymbol{\theta}_0\right) \\ \vdots \\ \mu_{n,k}\left(\boldsymbol{\theta}_0\right) \end{bmatrix}, \quad 1 \leq k \leq K-1.$$

Define $\tilde{\mathbf{Z}}_{\mathcal{S}}$ as a $(K-1) \times (K-1)$ block matrix,

$$\tilde{\mathbf{Z}}_{\mathcal{S}} = \begin{bmatrix} \mathbf{Z}_{\mathcal{S}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{\mathcal{S}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_{\mathcal{S}} \end{bmatrix}, \tag{29}$$

and $\mathbf{W}$ is an $n\left(K-1\right) \times n\left(K-1\right)$ square matrix,

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{1,1} & \mathbf{W}_{1,2} & \cdots & \mathbf{W}_{1,(K-1)} \\ \mathbf{W}_{2,1} & \mathbf{W}_{2,2} & \cdots & \mathbf{W}_{2,(K-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{W}_{(K-1),1} & \mathbf{W}_{(K-1),2} & \cdots & \mathbf{W}_{(K-1),(K-1)} \end{bmatrix}, \tag{30}$$

where each $\mathbf{W}_{k_1,k_2}$, $1 \leq k_1, k_2 \leq K-1$, is an $n \times n$ diagonal matrix. If $k_1 = k_2 = k$, the ith diagonal element in $\mathbf{W}_{k,k}$ is $\mu_{i,k}(\boldsymbol{\theta}_0)(1-\mu_{i,k}(\boldsymbol{\theta}_0))$. If $k_1 \neq k_2$, the ith diagonal element in $\mathbf{W}_{k,k}$ is $-\mu_{i,k_1}(\boldsymbol{\theta}_0)\mu_{i,k_2}(\boldsymbol{\theta}_0)$. Then

$$\boldsymbol{s}_{\mathcal{S}}(\boldsymbol{\theta}_0) = \tilde{\mathbf{Z}}_{\mathcal{S}}^T(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}}), \quad \mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_0) = \tilde{\mathbf{Z}}_{\mathcal{S}}^T \mathbf{W} \tilde{\mathbf{Z}}_{\mathcal{S}}. \tag{31}$$

Let $\boldsymbol{u} \in \mathbb{R}^{|\mathcal{S}|(K-1)}$ be a unit vector. We first show that

$$\begin{aligned} & \Pr\left\{\boldsymbol{u}^T [\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_0)]^{-1/2} \boldsymbol{s}_{\mathcal{S}}(\boldsymbol{\theta}_0) \geq \sqrt{2[|\mathcal{S}| - |\mathcal{A}|]\log(n^{r_1}q^{1+r_2})}\right\} \\ \leq & \exp\left\{-[|\mathcal{S}| - |\mathcal{A}|]\log\left(n^{r_1}q^{1+r_2}\right)\left(1 - \sqrt{\frac{Q\log(n^{r_1}q^{1+r_2})}{\lambda_1^3\lambda_3^{-2}n}}\right)\right\}. \end{aligned}$$

Define $A = \sqrt{2[|\mathcal{S}| - |\mathcal{A}|]\log(n^{r_1}q^{1+r_2})}$ and vector $\boldsymbol{\psi} \in \mathbb{R}^{|\mathcal{S}|(K-1)}$ with $\boldsymbol{\psi}_{\mathcal{S}} = A \cdot [\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_0)]^{-1/2}\boldsymbol{u}$ and other elements set as zero. We know by definition $\mathrm{Var}(\boldsymbol{s}_{\mathcal{S}}(\boldsymbol{\theta}_0)) = \mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_0)$, and

$$\left\|\mathbf{W}^{1/2}\tilde{\mathbf{Z}}_{\mathcal{S}}\boldsymbol{\psi}_{\mathcal{S}}\right\|_2^2 = A^2 \cdot \boldsymbol{u}^T [\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_0)]^{-1/2} \tilde{\mathbf{Z}}_{\mathcal{S}}^T \mathbf{W} \tilde{\mathbf{Z}}_{\mathcal{S}} [\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_0)]^{-1/2} \boldsymbol{u} = A^2.$$

and by Lemma 2, with probability at least $1 - C_1 n^{2\kappa - M} - C_2 n^{2\kappa - m} - C_3 n^{2\kappa Q - m}$, for all $\mathcal{S}$ with $|\mathcal{S}| \leq Q$,

$$\|\boldsymbol{\psi}\|_2^2 = \|\boldsymbol{\psi}_{\mathcal{S}}\|_2^2 = A^2 \cdot \left\|[\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_0)]^{-1/2}\boldsymbol{u}\right\|_2^2 \leq A^2 \cdot \left\|[\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_0)]^{-1}\right\|_{\text{sp}} \|\boldsymbol{u}\|_2^2 \leq A^2(\lambda_1 n)^{-1}.$$

We then have

$$\begin{aligned} & \Pr\left\{\boldsymbol{u}^T [\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_0)]^{-1/2} \boldsymbol{s}_{\mathcal{S}}(\boldsymbol{\theta}_0) \geq A\right\} \\ = & \mathbb{E}\left\{\mathbb{I}\left\{\boldsymbol{\psi}_{\mathcal{S}}^T \tilde{\mathbf{Z}}_{\mathcal{S}}^T(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}}) \geq A^2\right\}\right\} \\ = & \mathbb{E}\left\{\exp\left\{\boldsymbol{\psi}_{\mathcal{S}}^T \tilde{\mathbf{Z}}_{\mathcal{S}}^T(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}}) - A^2\right\}\right\} \\ = & \exp\left\{-\boldsymbol{\psi}_{\mathcal{S}}^T \tilde{\mathbf{Z}}_{\mathcal{S}}^T \tilde{\boldsymbol{\mu}} - A^2\right\} \cdot \mathbb{E}\left\{\exp\left\{\boldsymbol{\psi}_{\mathcal{S}}^T \tilde{\mathbf{Z}}_{\mathcal{S}}^T \tilde{\mathbf{y}}\right\}\right\} \\ = & \exp\left\{-\sum_{i=1}^n \boldsymbol{\psi}_{\mathcal{S}}^T \tilde{\mathbf{Z}}_{i,\mathcal{S}}^T \tilde{\boldsymbol{\mu}}_i - A^2\right\} \cdot \mathbb{E}\left\{\exp\left\{\sum_{i=1}^n \boldsymbol{\psi}_{\mathcal{S}}^T \tilde{\mathbf{Z}}_{i,\mathcal{S}}^T \tilde{\mathbf{y}}_i\right\}\right\}, \end{aligned} \tag{32}$$

where

$$\tilde{\mathbf{y}}_i = \begin{bmatrix} \mathbb{I}\{y_i = 1\} \\ \mathbb{I}\{y_i = 2\} \\ \vdots \\ \mathbb{I}\{y_i = K-1\} \end{bmatrix}, \quad \tilde{\boldsymbol{\mu}}_i = \begin{bmatrix} \mu_{i,1}(\boldsymbol{\theta}_0) \\ \mu_{i,2}(\boldsymbol{\theta}_0) \\ \vdots \\ \mu_{i,K-1}(\boldsymbol{\theta}_0) \end{bmatrix}, \quad \tilde{\mathbf{Z}}_{i,\mathcal{S}} = \begin{bmatrix} \mathbf{Z}_{i,\mathcal{S}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{i,\mathcal{S}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_{i,\mathcal{S}} \end{bmatrix}.$$

As $\tilde{\mathbf{y}}_i$ follows the vector exponential family with cumulant generating function

$$b\left(\boldsymbol{\theta}_{0,1}^T\mathbf{z}_i,\ldots,\boldsymbol{\theta}_{0,K-1}^T\mathbf{z}_i\right)=\log\left(1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{0,l}^T\mathbf{z}_i\right)\right),$$

with

$$\frac{\partial b\left(\boldsymbol{\theta}_{0,1}^T\mathbf{z}_i,\ldots,\boldsymbol{\theta}_{0,K-1}^T\mathbf{z}_i\right)}{\partial\left(\boldsymbol{\theta}_{0,k}^T\mathbf{z}_i\right)}=\frac{\exp\left(\boldsymbol{\theta}_{0,k}^T\mathbf{z}_i\right)}{1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{0,l}^T\mathbf{z}_i\right)}=\mu_{i,k}\left(\boldsymbol{\theta}_0\right).$$

By the property of exponential family,

$$\begin{aligned}
&\mathbb{E}\left\{\exp\left\{\sum_{i=1}^n\boldsymbol{\psi}_{\mathcal{S}}^T\tilde{\mathbf{Z}}_{i,\mathcal{S}}^T\tilde{\mathbf{y}}_i\right\}\right\}\\
=&\exp\left\{\sum_{i=1}^n\log\left(1+\sum_{l=1}^{K-1}\exp\left(\left(\left(\boldsymbol{\theta}_0+\boldsymbol{\psi}\right)_l^T\right)\mathbf{z}_i\right)\right)-\log\left(1+\sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{0,l}^T\mathbf{z}_i\right)\right)\right\}\\
=&\exp\left\{\sum_{i=1}^n b\left(\left(\left(\boldsymbol{\theta}_0+\boldsymbol{\psi}\right)_1^T\right)\mathbf{z}_i,\cdots,\left(\left(\boldsymbol{\theta}_0+\boldsymbol{\psi}\right)_{K-1}^T\right)\mathbf{z}_i\right)-b\left(\boldsymbol{\theta}_{0,1}^T\mathbf{z}_i,\cdots,\boldsymbol{\theta}_{0,K-1}^T\mathbf{z}_i\right)\right\},
\end{aligned}$$

and there exists $t\in(0,1)$,

$$\begin{aligned}
&\sum_{i=1}^n b\left(\left(\left(\boldsymbol{\theta}_0+\boldsymbol{\psi}\right)_1^T\right)\mathbf{z}_{i,\mathcal{S}},\cdots,\left(\left(\boldsymbol{\theta}_0+\boldsymbol{\psi}\right)_{K-1}^T\right)\mathbf{z}_{i,\mathcal{S}}\right)-b\left(\boldsymbol{\theta}_{0,1}^T\mathbf{z}_i,\cdots,\boldsymbol{\theta}_{0,K-1}^T\mathbf{z}_i\right)\\
=&\sum_{i=1}^n\left(\boldsymbol{\psi}_1^T\mathbf{z}_i,\ldots,\boldsymbol{\psi}_{K-1}^T\mathbf{z}_i\right)^T\nabla b\left(\boldsymbol{\theta}_{0,1}^T\mathbf{z}_i,\cdots,\boldsymbol{\theta}_{0,K-1}^T\mathbf{z}_i\right)\\
&+\frac{1}{2}\left(\boldsymbol{\psi}_1^T\mathbf{z}_i,\ldots,\boldsymbol{\psi}_{K-1}^T\mathbf{z}_i\right)^T\nabla^2 b\left(\left(\boldsymbol{\theta}_0+t\boldsymbol{\psi}\right)_1^T\mathbf{z}_i,\cdots,\left(\boldsymbol{\theta}_0+t\boldsymbol{\psi}\right)_{K-1}^T\mathbf{z}_i\right)\left(\boldsymbol{\psi}_1^T\mathbf{z}_i,\ldots,\boldsymbol{\psi}_{K-1}^T\mathbf{z}_i\right)\\
=&\left[\sum_{i=1}^n\boldsymbol{\psi}_{\mathcal{S}}^T\tilde{\mathbf{Z}}_{i,\mathcal{S}}^T\tilde{\boldsymbol{\mu}}_i\right]+\frac{1}{2}\boldsymbol{\psi}_{\mathcal{S}}^T\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right)\boldsymbol{\psi}_{\mathcal{S}}+\frac{1}{2}\boldsymbol{\psi}_{\mathcal{S}}^T\left[\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{0,\mathcal{S}}+t\boldsymbol{\psi}_{\mathcal{S}}\right)-\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_{0,\mathcal{S}}\right)\right]\boldsymbol{\psi}_{\mathcal{S}}\\
\leq&\left[\sum_{i=1}^n\boldsymbol{\psi}_{\mathcal{S}}^T\tilde{\mathbf{Z}}_{i,\mathcal{S}}^T\tilde{\boldsymbol{\mu}}_i\right]+\frac{A^2}{2}+\frac{1}{2}\left\|\boldsymbol{\psi}_{\mathcal{S}}\right\|_2^3 n\lambda_3\\
\leq&\left[\sum_{i=1}^n\boldsymbol{\psi}_{\mathcal{S}}^T\tilde{\mathbf{Z}}_{i,\mathcal{S}}^T\tilde{\boldsymbol{\mu}}_i\right]+\frac{A^2}{2}+\frac{A^3}{2}n^{-0.5}\left(\lambda_1\right)^{-3/2}\lambda_3.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\Pr\left\{\boldsymbol{u}^T\left[\mathbf{H}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right)\right]^{-1/2}\boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right)\geq A\right\}\\
\leq&\exp\left\{-\frac{A^2}{2}+\frac{A^3}{2}n^{-0.5}\left(\lambda_1\right)^{-3/2}\lambda_3\right\},\\
=&\exp\left\{-\left[|\mathcal{S}|-|\mathcal{A}|\right]\log\left(n^{r_1}q^{1+r_2}\right)\left(1-\sqrt{\frac{2Q\log\left(n^{r_1}q^{1+r_2}\right)}{\lambda_1^3\lambda_3^{-2}n}}\right)\right\}.
\end{aligned}$$

In next step, following the approach in the proof of Foygel and Drton (2011) Lemma (i), the

following inequality can be shown to hold. The details are omitted for brevity.

$$\Pr\left\{\exists \mathcal{S},\, \mathcal{S} \supsetneq \mathcal{A},\, |\mathcal{S}| \leq Q,\, \boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right)^T \mathbf{H}_{\mathcal{S}}^{-1}\left(\boldsymbol{\theta}_0\right) \boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right) \geq 2\left[|\mathcal{S}| - |\mathcal{A}|\right] \log\left(n^{r_1} q^{1+r_2}\right)\right\} \leq \frac{1}{3} n^{-r_1} q^{-r_2}.$$

Therefore there exists constants $r_1, r_2 > 0$, with probability at least $1 - n^{-r_1} q^{-r_2}$, uniformly for all $\mathcal{S} \supsetneq \mathcal{A}$ with $|\mathcal{S}| \leq Q$,

$$l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right) - l_n\left(\boldsymbol{\theta}_0\right) \leq \left[|\mathcal{S}| - |\mathcal{A}|\right] \log\left(n^{r_1} q^{1+r_2}\right) + \varepsilon_n.$$

2.2 Proof of Theorem 1

Theorem 1. *Under conditions C1 ~ C4, as $n \to \infty$,*

$$\max_{\mathcal{S} \supset \mathcal{A},\, |\mathcal{S}| \leq Q} \left\| \tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0 \right\|_2 = O_p\left(n^{-1/2+\xi}\right), \tag{33}$$

where $0 < \xi < 1/2$ and $Q \geq |\mathcal{A}|$ are any positive constants independent of n.

Proof. There exists $t \in (0,1)$ such that

$$\begin{aligned} l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right) - l_n\left(\boldsymbol{\theta}_0\right) &= \left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)^T \boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right) - \frac{1}{2}\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)^T \mathbf{H}_{\mathcal{S}}\left(t\tilde{\boldsymbol{\theta}}_{\mathcal{S}} + (1-t)\,\boldsymbol{\theta}_0\right)\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right) \\ &\leq \left\| \tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0 \right\|_2 \left\| \boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right) \right\|_2 - \frac{\lambda_1 n}{2} \left\| \tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0 \right\|_2^2, \end{aligned}$$

with probability tending to 1, uniformly for all $\mathcal{S} \supset \mathcal{A}, |\mathcal{S}| \leq Q$, as $n \to \infty$ by Lemma 2. Since vector $\boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right)$ has only finite number of elements, $\left\| \boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right) \right\|_2 = O\left(\left\| \boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right) \right\|_\infty\right)$. We first calculate the following uniform bound for $\left\| \boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right) \right\|_\infty$.

Let ξ be any constant such that $0 < \xi < 1/2$

$$\begin{aligned} &\Pr\left(\max_{\mathcal{S} \supset \mathcal{A},\, |\mathcal{S}| \leq Q} \left\| \boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right) \right\|_\infty \geq C_1 n^{1/2+\xi}\right) \\ \leq &\sum_{\mathcal{S} \supset \mathcal{A},\, |\mathcal{S}| \leq Q} \sum_{k=1}^{K-1} \sum_{j \in \mathcal{S}} \left[P\left(\left| s_{k,j}\left(\boldsymbol{\theta}_0\right) \right| \geq C_2 n^{1/2+\xi}\right)\right], \end{aligned}$$

where constants $C_1, C_2 > 0$ and

$$s_{k,j}\left(\boldsymbol{\theta}_0\right) = \sum_{i=1}^{n} \left(\mathbb{I}\left\{y_i = k\right\} - \mu_{i,k}\left(\boldsymbol{\theta}_0\right)\right) z_{i,j}.$$

By definition of $\boldsymbol{\theta}_0$,

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}_{\mathcal{S}}:\, \mathcal{S} \supset \mathcal{A}}{\arg\max}\, \mathbb{E}\left[\log p\left(Y \mid \mathbf{Z}, \boldsymbol{\theta}_{\mathcal{S}}\right)\right].$$

Therefore for all $j \in \mathcal{S}$ and $k = 1, \ldots, K-1$,

$$\mathbb{E}\left[\frac{\partial \log p\left(Y \mid \mathbf{Z}, \boldsymbol{\theta}\right)}{\partial \theta_{k,j}}\right] = \mathbb{E}\left\{\left[\mathbb{I}\left\{Y = k\right\} - \mu_k\left(\boldsymbol{\theta}_0\right)\right] Z_j\right\} = 0,$$

where

$$\mu_k\left(\boldsymbol{\theta}_0\right) = \frac{\exp\left(\boldsymbol{\theta}_0 \mathbf{Z}\right)}{1 + \sum_{l=1}^{K-1} \exp\left(\boldsymbol{\theta}_0 \mathbf{Z}\right)}$$

where Y and $\mathbf{Z}$ are random variables, and the expectation is taken over sampling distribution of $(\mathbf{Z}, Y)$.

By condition C3, each Z_j is sub-exponential. Since $\left(y_{i,k} - \mu_{i,k}\left(\boldsymbol{\theta}_0\right)\right) \in (-1, 1)$, $\left(y_{i,k} - \mu_{i,k}\left(\boldsymbol{\theta}_0\right)\right) z_{i,j}$ is also sub-exponential. By Bernstein's inequality, there exist constants C_3, C_4, C_5 such that

$$\begin{aligned}
\Pr\left(\left|s_{k,j}\left(\boldsymbol{\theta}_0\right)\right| > n\varepsilon\right) &= \Pr\left(\left|\sum_{i=1}^{n}\left(y_{i,k} - \mu_{i,k}\left(\boldsymbol{\theta}_0\right)\right) z_{i,j}\right| > n\varepsilon\right) \\
&\leq C_3 \exp\left(-C_4 n \varepsilon^2\right), \quad \text{for } |\varepsilon| \leq C_5.
\end{aligned}$$

Let $\varepsilon = C_2 n^{-1/2+\xi}$, then

$$\Pr\left(\left|s_{k,j}\left(\boldsymbol{\theta}_0\right)\right| > C_2 n^{1/2+\xi}\right) \leq C_3 \exp\left(-C_2^2 C_4 n^{2\xi}\right).$$

Since $q = \frac{1}{2}p\left(p+3\right) \leq p^2$ for $p \geq 3$, $q = O\left(n^{2\kappa}\right)$,

$$\begin{aligned}
\Pr\left(\max_{\mathcal{S} \supset \mathcal{A},\, |\mathcal{S}| \leq Q} \left\|\boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right)\right\|_\infty \geq C_1 n\varepsilon\right) &\leq \sum_{\mathcal{S} \supset \mathcal{A},\, |\mathcal{S}| \leq Q} \sum_{k=1}^{K-1} \sum_{j \in \mathcal{S}} \left[\Pr\left(\left|s_{k,j}\left(\boldsymbol{\theta}_0\right)\right| \geq C_2 n^{1/2+\xi}\right)\right] \\
&\leq q^Q \cdot (K-1) \cdot Q \cdot C_3 \exp\left(-C_2^2 C_4 n^{2\xi}\right) \\
&\leq (K-1) \cdot Q \cdot C_3 \exp\left(-C_2^2 C_4 n^{2\xi} + 2\kappa Q \log n\right) \to 0,
\end{aligned}$$

as $n \to \infty$. Recall that,

$$l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right) - l_n\left(\boldsymbol{\theta}_0\right) \leq \left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right\|_2 \left\|\boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right)\right\|_2 - \frac{\lambda_1 n}{2}\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right\|_2^2.$$

Because of l_n's concavity and $\mathcal{S} \supset \mathcal{A}$, $l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right) - l_n\left(\boldsymbol{\theta}_0\right) \geq 0$. Therefore we must have

$$\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right\|_2 \left\|\boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right)\right\|_2 - \frac{\lambda_1 n}{2}\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right\|_2^2 \geq 0, \tag{34}$$

which implies

$$\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right\|_2 \leq \frac{2}{\lambda_1 n}\left\|\boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right)\right\|_2. \tag{35}$$

Since $\|\boldsymbol{s}_{\mathcal{S}}(\boldsymbol{\theta}_0)\|_2 = O_p\left(n^{1/2+\xi}\right)$, then as $n \to \infty$,

$$\max_{\mathcal{S}\supset\mathcal{A},\,|\mathcal{S}|\leq Q} \left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right\|_2 = O_p\left(n^{-1/2+\xi}\right).$$

2.3 Proof of Theorem 3

Theorem 2. ***(Forward stage screening consistency)*** *If conditions C1 ~ C4 hold, and all predictors in $\mathcal{P}$ are stepwise detectable, then the forward interaction screening stage finishes in finite number of steps and is screening consistent. In particular, as $n \to \infty$,*

$$Pr\left(\left|\tilde{\mathcal{C}}_F\right| \leq Q\right) \to 1, \quad and \quad Pr\left(\tilde{\mathcal{C}}_F \supseteq \mathcal{P}\right) \to 1,$$

where $Q = \left\lceil 8\lambda_1^{-1}\theta_{\min}^{-2}\log K\right\rceil$, λ_1 is a positive constant defined in Lemma 2 and $\theta_{\min}$ is a positive constant defined in condition C2.

Proof. The proof consists of two parts. In the first part we will show that

$$P\left(\tilde{\mathcal{C}}_F \supseteq \mathcal{P}\right) \to 1. \tag{36}$$

In the second part, we will show that with probability going to 1, $\left|\tilde{\mathcal{C}}_F\right| \leq Q$, therefore in the first part of the proof we will only consider sets $\mathcal{C}$ that $|\mathcal{C}| \leq Q$.

Part I: Let $\mathcal{C}$ denote a set of predictors, and let $\mathcal{S}_{\mathcal{C}}$ denote the set of corresponding terms in forward interaction screening stage. By definition, $\mathcal{S}_{\mathcal{C}} = \mathcal{C} \cup (\mathcal{C} \times \mathcal{C})$ and contains all main effect and interaction terms for predictors in $\mathcal{C}$. We will show the uniform bound,

$$\Pr\left(\max_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{P}\neq\emptyset,|\mathcal{C}|\leq Q}\min_{j\in\mathcal{C}^c}\left\{\text{EBIC}_\gamma\left(\mathcal{S}_{\mathcal{C}\cup\{j\}}\right) - \text{EBIC}_\gamma\left(\mathcal{S}_{\mathcal{C}}\right)\right\} < 0\right) \to 1. \tag{37}$$

This implies that if current set $\mathcal{C}$ does not contain all predictors in $\mathcal{P}$, then, with probability tending to 1, we can always find a new predictor $j \in \mathcal{C}^c$ such that $\text{EBIC}_\gamma\left(\mathcal{S}_{\mathcal{C}\cup\{j\}}\right) - \text{EBIC}_\gamma\left(\mathcal{S}_{\mathcal{C}}\right) < 0$. Therefore, forward screening stage will proceed until all predictors are added into the model, and eventually stop at some $\tilde{\mathcal{C}}_F \supseteq \mathcal{P}$, which also implies $\mathcal{S}_{\tilde{\mathcal{C}}_F} \supseteq \mathcal{A}$.

When $\mathcal{C}^c \cap \mathcal{P} \neq \emptyset$ and all the relevant predictors in $\mathcal{P}$ are stepwise detectable, there exists $m \geq 0$ such that $\cup_{i=0}^{m-1}\mathcal{T}_i \subset \mathcal{C}$ and $\mathcal{C}^c \cap \mathcal{T}_m \neq \emptyset$. According to definition of stepwise detectable condition, there exists $j \in \mathcal{C}^c \cap \mathcal{T}_m$ and constants $\theta_{\max} > \theta_{\min} > 0$ such that

$$\theta_{\min} \leq \left\|\boldsymbol{\theta}^{j*}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}}\right\|_\infty \leq \theta_{\max}.$$

Vector $\boldsymbol{\theta}^{j*}_{\mathcal{C}^j}$ contains the parameters in $\boldsymbol{\theta}^*_{\mathcal{C}^j}$ associated with predictor X_j. By the Mean Value Theo-

rem, there exists $t \in (0,1)$,

$$\begin{aligned}
& \max_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{P}\neq\emptyset,|\mathcal{C}|\leq Q} \min_{j\in\mathcal{C}^c} \left[l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}}\right) - l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}}\right)\right] \\
\leq & \max_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{P}\neq\emptyset,|\mathcal{C}|\leq Q} \min_{j\in\mathcal{C}^c\cap\mathcal{T}_m} \left[l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}}\right) - l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}}\right)\right] \\
= & \max_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{P}\neq\emptyset,|\mathcal{C}|\leq Q} \min_{j\in\mathcal{C}^c\cap\mathcal{T}_m} \left[-\frac{1}{2}\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}}\right)^T \mathbf{H}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}}\left(t\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} + (1-t)\,\tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}}\right)\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}}\right)\right] \\
\leq & \max_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{P}\neq\emptyset,|\mathcal{C}|\leq Q} \min_{j\in\mathcal{C}^c\cap\mathcal{T}_m} \left[-\frac{\lambda_1 n}{2}\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}}\right\|_2^2\right].
\end{aligned}$$

with probability going to 1 for all $|\mathcal{C}| \leq Q$ by Lemma 2. We then give a lower bound to $\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}}\right\|_2^2$.

We first consider $\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right\|_2^2$ and $\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}} - \boldsymbol{\theta}^*_{\mathcal{S}_{\mathcal{C}\cup\{j\}}}\right\|_2^2$. There exists $t \in (0,1)$,

$$\begin{aligned}
l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}}\right) - l_n\left(\boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right) &= \left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right)^T \boldsymbol{s}_\mathcal{S}\left(\boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right) - \frac{1}{2}\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right)^T \mathbf{H}_\mathcal{S}\left(t\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} + (1-t)\,\boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right)\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right) \\
&\leq \left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right\|_2 \left\|\boldsymbol{s}_\mathcal{S}\left(\boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right)\right\|_2 - \frac{\lambda_1 n}{2}\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right\|_2^2,
\end{aligned}$$

uniformly for all $\mathcal{C}$ such that $|\mathcal{C}| \leq Q$, as $n \to \infty$, with probability tending to $1 - C_1 n^{2\kappa-M} - C_2 n^{2\kappa-m} - C_3 n^{2\kappa Q-m} \to 1$ by Lemma 2.

Because vector $\boldsymbol{s}_\mathcal{S}\left(\boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right)$ has only finite number of elements, $\left\|\boldsymbol{s}_\mathcal{S}\left(\boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right)\right\|$ is in the same order of n as $\left\|\boldsymbol{s}_\mathcal{S}\left(\boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right)\right\|_\infty$. We first calculate the following uniform bound for $\left\|\boldsymbol{s}_\mathcal{S}\left(\boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right)\right\|_\infty$. Using the method in proof of theorem 1, we get the following uniform bound for all $|\mathcal{C}| \leq Q$. For any $\xi > 0$, there exists a constant $C_1 > 0$,

$$\Pr\left(\max_{\mathcal{C}:|\mathcal{C}|\leq Q}\left\|\boldsymbol{s}_\mathcal{S}\left(\boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right)\right\|_\infty \geq C_1 n^{1/2+\xi}\right) \quad \to \quad 0,$$

as $n \to \infty$. So $\left\|\boldsymbol{s}_\mathcal{S}\left(\boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right)\right\|_2 = O_p\left(n^{1/2+\xi}\right)$, and

$$l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}}\right) - l_n\left(\boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right) \quad \leq \quad \left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right\|_2 O_p\left(n^{1/2+\xi}\right) - \frac{\lambda_1 n}{2}\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right\|_2^2.$$

Because of l_n's concavity, $l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}}\right) - l_n\left(\boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right) > 0$. Therefore we must have

$$\max_{\mathcal{C}:|\mathcal{C}|\leq Q}\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \boldsymbol{\theta}^*_{\mathcal{S}_\mathcal{C}}\right\|_2 = O_p\left(n^{-1/2+\xi}\right),$$

as $n \to \infty$. Similarly

$$\max_{\mathcal{C}:|\mathcal{C}|\leq Q}\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}} - \boldsymbol{\theta}^*_{\mathcal{S}_{\mathcal{C}\cup\{j\}}}\right\|_2 = O_p\left(n^{-1/2+\xi}\right),$$

Now we give lower bound to $\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}}\right\|_2^2$. By the definition of stepwise detectable condi-

tion, there exists $j \in \mathcal{S}^c \cap \mathcal{T}_m$ such that

$$\theta_{\min} \leq \left\| \boldsymbol{\theta}^{j*}_{\mathcal{S}\cup\{j\}} \right\|_\infty \leq \theta_{\max}.$$

Therefore with probability tending to 1 as $n \to \infty$, uniformly for all $\mathcal{C}$ such that $\mathcal{C}^c \cap \mathcal{P} \neq \emptyset$ and $|\mathcal{C}| \leq Q$,

$$\begin{aligned}
\max_{j\in\mathcal{C}^c} \left\| \tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}} \right\|_2^2 &\geq \max_{j\in\mathcal{C}^c\cap\mathcal{T}_m} \left\| \tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}} \right\|^2 \\
&\geq \max_{j\in\mathcal{C}^c\cap\mathcal{T}_m} \sum_{k=1}^{K-1} \sum_{l\in\mathcal{S}_{\mathcal{C}\cup\{j\}}\setminus\mathcal{S}_\mathcal{C}} \tilde{\theta}^2_{l,k,\mathcal{S}_{\mathcal{C}\cup\{j\}}} \\
&\geq \max_{j\in\mathcal{C}^c\cap\mathcal{T}_m} \max_{k=1,\ldots,K-1} \max_{l\in\mathcal{S}_{\mathcal{C}\cup\{j\}}\setminus\mathcal{S}_\mathcal{C}} \tilde{\theta}^2_{l,k,\mathcal{S}_{\mathcal{C}\cup\{j\}}} \\
&\geq \max_{j\in\mathcal{C}^c\cap\mathcal{T}_m} \max_{k=1,\ldots,K-1} \max_{l\in\mathcal{S}_{\mathcal{C}\cup\{j\}}\setminus\mathcal{S}_\mathcal{C}} \left(\theta^*_{l,k,\mathcal{S}_{\mathcal{C}\cup\{j\}}} + \tilde{\theta}_{l,k,\mathcal{S}_{\mathcal{C}\cup\{j\}}} - \theta^*_{l,k,\mathcal{S}_{\mathcal{C}\cup\{j\}}} \right)^2 \\
&= \max_{j\in\mathcal{C}^c\cap\mathcal{T}_m} \max_{k=1,\ldots,K-1} \max_{l\in\mathcal{S}_{\mathcal{C}\cup\{j\}}\setminus\mathcal{S}_\mathcal{C}} \left(\theta^{*2}_{l,k,\mathcal{S}_{\mathcal{C}\cup\{j\}}} + O_p\left(n^{-1/2+\xi}\right) \right), \\
&\geq \frac{1}{2}\theta^2_{\min},
\end{aligned}$$

as $n \to \infty$. Eventually, we have

$$\begin{aligned}
& \max_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{P}\neq\emptyset,|\mathcal{C}|\leq Q} \min_{j\in\mathcal{C}^c} \left[l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}}\right) - l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}}\right) \right] \\
\leq & \max_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{P}\neq\emptyset,|\mathcal{C}|\leq Q} \min_{j\in\mathcal{C}^c\cap\mathcal{T}_m} \left[l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}}\right) - l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}}\right) \right] \\
\leq & \max_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{P}\neq\emptyset,|\mathcal{C}|\leq Q} \min_{j\in\mathcal{C}^c\cap\mathcal{T}_m} \left[-\frac{\lambda_1 n}{2} \left\| \tilde{\boldsymbol{\theta}}_{\mathcal{S}_\mathcal{C}} - \tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}} \right\|^2 \right] \\
\leq & -\frac{1}{4} n\lambda_1\theta^2_{\min}. \qquad (38)
\end{aligned}$$

Therefore, as $n \to \infty$,

$$\begin{aligned}
\text{EBIC}_\gamma\left(\mathcal{S}_{\mathcal{C}\cup\{j\}}\right) - \text{EBIC}_\gamma\left(\mathcal{S}_\mathcal{C}\right) &\leq -\frac{1}{4}n\lambda_1\theta^2_{\min} + \left[\left|\mathcal{S}_{\mathcal{C}\cup\{j\}}\right| - |\mathcal{S}_\mathcal{C}| \right] \left(\frac{1}{2}\log n + \gamma \log p \right) \\
&\leq -n\left(\frac{1}{4}\lambda_1\theta^2_{\min} - \frac{1}{n}\left[\left|\mathcal{S}_{\mathcal{C}\cup\{j\}}\right| - |\mathcal{S}_\mathcal{C}| \right] \left(\frac{1}{2}\log n + \gamma\log p \right) \right) \\
&< 0,
\end{aligned}$$

holds uniformly for all $|\mathcal{C}| \leq Q$ and $j \in \mathcal{A}$ with probability going to 1, and

$$P\left(\max_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{P}\neq\emptyset,|\mathcal{C}|\leq Q} \min_{j\in\mathcal{C}^c} \left\{ \text{EBIC}_\gamma\left(\mathcal{S}_{\mathcal{C}\cup\{j\}}\right) - \text{EBIC}_\gamma\left(\mathcal{S}_\mathcal{C}\right) \right\} < 0 \right) \to 1, \qquad (39)$$

thus we proved that $P\left(\tilde{\mathcal{S}}_F \supseteq \mathcal{A}\right) \to 1$, as $n \to \infty$.

Part II: In this part, we will show that as $n \to \infty$, with probability tending to 1, the forward

interaction screening stage will stop in a finite number of steps. In particular, we will show that the number of steps in forward stage cannot exceed $Q = \left\lceil 8\lambda_1^{-1}\theta_{\min}^{-2}\log K \right\rceil$.

Let $\mathcal{C}_1, \mathcal{C}_2, \ldots, \tilde{\mathcal{C}}_F$ denote the selected set of predictors in each step of the forward interaction screening stage. By definition $\mathcal{C}_1 = \emptyset$, so $\Pr\left(Y = k \mid \mathbf{X}, \boldsymbol{\theta}_{\mathcal{C}_1}\right) = 1/K$ for $k = 1, \ldots, K$, and

$$l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}_1}}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{I}\left\{y_i = k\right\}\log\left(1/K\right) = n\log\left(1/K\right).$$

Define $G_n\left(\mathcal{C}\right) = -\frac{1}{n}l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}}}\right)$, then

$$G_n\left(\mathcal{C}_1\right) = -\frac{1}{n}n\log\left(1/K\right) = \log K \tag{40}$$

By the nature of forward screening stage, $\mathcal{C}_1 \subset \mathcal{C}_2 \subset \cdots \subset \tilde{\mathcal{C}}_F$, thus

$$G_n\left(\mathcal{C}_1\right) > G_n\left(\mathcal{C}_2\right) > \cdots > G_n\left(\tilde{\mathcal{C}}_F\right) \geq 0.$$

Consider two adjacent sets $\mathcal{C}_m$ and $\mathcal{C}_{m+1}$. Uniformly for all m with $\mathcal{C}_m^c \cap \mathcal{P} \neq \emptyset, |\mathcal{C}_{m+1}| \leq Q$,

$$\begin{aligned}
G_n\left(\mathcal{C}_m\right) - G_n\left(\mathcal{C}_{m+1}\right) &= \frac{1}{n}\max_{j\in\mathcal{C}_m^c}\left[l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}_m\cup\{j\}}}\right) - l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}_m}}\right)\right] \\
&\geq \frac{1}{n}\max_{j\in\mathcal{C}_m^c\cap\mathcal{T}_m}\left[l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}_m\cup\{j\}}}\right) - l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}_{\mathcal{C}_m}}\right)\right] \\
&\geq \frac{1}{n}\left[\frac{1}{4}n\lambda_1\theta_{\min}^2\right] \\
&= \frac{1}{4}\lambda_1\theta_{\min}^2,
\end{aligned}$$

with probability going to 1 as $n \to \infty$, where $\lambda_1 > 0$, $\theta_{\min} > 0$ are constants defined earlier. Define $Q = \left\lceil 8\lambda_1^{-1}\theta_{\min}^{-2}\log K \right\rceil$. It is straightforward to see that

$$\begin{aligned}
G_n\left(\mathcal{C}_1\right) &= G_n\left(\mathcal{C}_{Q+1}\right) + \sum_{i=1}^{Q}\left(G_n\left(\mathcal{C}_i\right) - G_n\left(\mathcal{C}_{i+1}\right)\right) \\
&\geq \sum_{i=1}^{Q}\left(G_n\left(\mathcal{C}_i\right) - G_n\left(\mathcal{C}_{i+1}\right)\right) \\
&\geq Q\frac{1}{4}\lambda_1\theta_{\min}^2 \\
&> \log K.
\end{aligned} \tag{41}$$

Eq (40) contradicts to the fact $G_n\left(\mathcal{C}_1\right) = \log K$. Therefore with probability going to 1, the forward screening stage adds all predictors in $\mathcal{P}$ in less than $Q = \left\lceil 8\lambda_1^{-1}\theta_{\min}^{-2}\log K \right\rceil$ steps, and it is also straightforward to show that once $\mathcal{C}_m \supset \mathcal{P}$, the forward interaction screening stops at step m. In summary, forward interaction screening finishes in Q steps and $\left|\tilde{\mathcal{C}}_F\right| \leq Q$.

2.4 Proof of Theorem 4

Theorem 3. (Uniform bound of EBIC in backward stage) *Fix any positive constant $Q > 0$. Under conditions C1 ~ C4, as $n \to \infty$,*

$$Pr\left(\max_{\mathcal{S}\supsetneq\mathcal{A}:|\mathcal{S}|\leq Q}\min_{j\in\mathcal{S}\setminus\mathcal{A}}\left\{EBIC_\gamma\left(\mathcal{S}\setminus\{j\}\right)-EBIC_\gamma\left(\mathcal{S}\right)\right\}<0\right)\to 1, \tag{42}$$

and

$$Pr\left(\min_{\mathcal{S}\supset\mathcal{A}:|\mathcal{S}|\leq Q}\min_{j\in\mathcal{A}}\left\{EBIC_\gamma\left(\mathcal{S}\setminus\{j\}\right)-EBIC_\gamma\left(\mathcal{S}\right)\right\}<0\right)\to 0, \tag{43}$$

for any constant $\gamma > Q - |\mathcal{A}| - (2\kappa)^{-1}$.

Proof. Eq (42) implies that if $\mathcal{S}\supsetneq\mathcal{A}$ and $|\mathcal{S}|\leq Q$, with probability tending to 1, there will be at least one irrelevant term $j\in\mathcal{S}\cap\mathcal{A}^c$ such that removing j from $\mathcal{S}$ leads to lower EBIC.

For $j\in\mathcal{S}\setminus\mathcal{A}$, we have $\mathcal{A}\subseteq\mathcal{S}\setminus\{j\}\subsetneq\mathcal{S}$, and $l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right)\geq l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}\setminus\{j\}}\right)\geq l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{A}}\right)$. By Lemma 3, as $n\to\infty$, there exists constants $r_1, r_2 > 0$, with probability at least $1-n^{-r_1}q^{-r_2}$, uniformly for all $\mathcal{S}\supsetneq\mathcal{A}$ with $|\mathcal{S}|\leq Q$,

$$\begin{aligned} l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right)-l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}\setminus\{j\}}\right) &\leq l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right)-l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{A}}\right)\\ &\leq l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right)-l_n\left(\boldsymbol{\theta}_0\right)\\ &\leq \left[|\mathcal{S}|-|\mathcal{A}|\right]\log\left(n^{r_1}q^{1+r_2}\right)+\varepsilon_n, \end{aligned}$$

where $\boldsymbol{\theta}_0$ is the true parameters, $q=p\left(p+3\right)/2$ and $\epsilon_n=O\left(n^{-1/3}\right)$. Let $\Delta=|\mathcal{S}|-|\mathcal{A}|$. For all $\mathcal{S}\supsetneq\mathcal{A}:|\mathcal{S}|\leq Q$ and any $j\in\mathcal{S}\setminus\mathcal{A}$, with probability at least $1-n^{-r_1}q^{-r_2}$,

$$\begin{aligned} &\text{EBIC}_\gamma\left(\mathcal{S}\setminus\{j\}\right)-\text{EBIC}_\gamma\left(\mathcal{S}\right)\\ =\;& 2\left[l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right)-l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}\setminus\{j\}}\right)\right]-\left(\log n+2\gamma\log p\right)\\ \leq\;& \Delta\left(2r_1\log n+2\left(1+r_2\right)\log q\right)-\left(\log n+2\gamma\log p\right)\\ \leq\;& \left(2r_1\Delta-1\right)\log n+\left(2\Delta+2\Delta r_2-2\gamma\right)\log p\\ \leq\;& \left(\frac{2r_1\Delta-1}{\kappa}+2\Delta+2\Delta r_2-2\gamma\right)\log p\\ <\;& 0, \end{aligned}$$

with $\gamma>\Delta+\Delta r_2+\frac{2r_1\Delta-1}{2\kappa}$. Since $\Delta\leq Q-|\mathcal{A}|$ and $r_1, r_2>0$ can be arbitrarily small, we can choose any $\gamma>Q-|\mathcal{A}|-(2\kappa)^{-1}$. Uniformly for all any $\mathcal{S}\supsetneq\mathcal{A}:|\mathcal{S}|\leq Q$ and any $j\in\mathcal{S}\setminus\mathcal{A}$, we have

$$\max_{\mathcal{S}\supsetneq\mathcal{A}:|\mathcal{S}|\leq Q}\min_{j\in\mathcal{S}\setminus\mathcal{A}}\left\{\text{EBIC}_\gamma\left(\mathcal{S}\setminus\{j\}\right)-\text{EBIC}_\gamma\left(\mathcal{S}\right)\right\}<0,$$

with probability at least $1-n^{-r_1}q^{-r_2}\to 1$ as $n\to\infty$.

Eq (43) implies that if $\mathcal{S}\supset\mathcal{A}$ and $|\mathcal{S}|\leq Q$, then with probability tending to 1, there is no any relevant term $j\in\mathcal{A}$ such that removing j from $\mathcal{S}$ leads to lower EBIC. Therefore is no relevant

term will be removed from $\mathcal{S}$ as $n \to \infty$.

There exists $t \in (0,1)$ such that

$$\begin{aligned} l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}\setminus\{j\}}\right) - l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right) &\leq l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}\setminus\{j\}}\right) - l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{A}}\right) \\ &\leq l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}\setminus\{j\}}\right) - l_n\left(\boldsymbol{\theta}_0\right) \\ &= \left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)^T \boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right) - \frac{1}{2}\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right)^T \mathbf{H}_{\mathcal{S}}\left(t\tilde{\boldsymbol{\theta}}_{\mathcal{S}} + (1-t)\,\boldsymbol{\theta}_0\right)\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right) \\ &\leq \left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right\|_2 \left\|\boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right)\right\|_2 - \frac{\lambda_1 n}{2}\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right\|_2^2, \end{aligned}$$

uniformly for all $|\mathcal{S}| \leq Q$ with probability tending to 1 by Lemma 2.

$$\begin{aligned} l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}\setminus\{j\}}\right) - l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right) &\leq l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right) - l_n\left(\boldsymbol{\theta}_0\right) \\ &\leq \left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right\|_2 \left\|\boldsymbol{s}_{\mathcal{S}}\left(\boldsymbol{\theta}_0\right)\right\|_2 - \frac{\lambda_1 n}{2}\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right\|_2^2 \\ &= \left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right\|_2 O_p\left(n^{1/2+\xi}\right) - \frac{\lambda_1 n}{2}\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right\|_2^2. \end{aligned} \tag{44}$$

By condition C2, there exists a constant $\theta_{\min} > 0$ such that $\left\|\tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0\right\|_2 \geq \theta_{\min}$. Thus with sufficiently large n, there exists positive constants C_1,

$$l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}\setminus\{j\}}\right) - l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right) \leq -C_1\theta_{\min}^2 n,$$

and as $n \to \infty$, with probability going to 1,

$$\text{EBIC}_\gamma\left(\mathcal{S}\setminus\{j\}\right) - \text{EBIC}_\gamma\left(\mathcal{S}\right) \geq 2C_1\theta_{\min}^2 n - \left[|\mathcal{S}| - |\mathcal{A}|\right]\left(\log n + 2\gamma\log p\right) > 0, \tag{45}$$

uniformly for all $\mathcal{S} \supset \mathcal{A} : |\mathcal{S}| \leq Q$ and $j \in \mathcal{A}$, which indicates that with probability going to 0,

$$\min_{\mathcal{S}\supset\mathcal{A}:|\mathcal{S}|\leq Q} \min_{j\in\mathcal{A}} \left\{\text{EBIC}_\gamma\left(\mathcal{S}\setminus\{j\}\right) - \text{EBIC}_\gamma\left(\mathcal{S}\right)\right\} < 0.$$

References

Beer, D. G., S. L. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature medicine 8*(8), 816–824.

Efron, B. (2009). Empirical bayes estimates for large-scale prediction problems. *Journal of the American Statistical Association 104*(487), 1015–1028.

Fan, Y., Y. Kong, D. Li, Z. Zheng, et al. (2015). Innovated interaction screening for high-dimensional nonlinear classification. *The Annals of Statistics 43*(3), 1243–1272.

Foygel, R. and M. Drton (2011). Bayesian model choice and information criteria in sparse generalized linear models. *arXiv preprint arXiv:1112.5635*.

Ravikumar, P., M. J. Wainwright, J. D. Lafferty, et al. (2010). High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics 38*(3), 1287–1319.

Ravikumar, P., M. J. Wainwright, G. Raskutti, and B. Yu (2011). High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electron. J. Statist. 5*, 935–980.

Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell 1*(2), 203–209.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences 99*(10), 6567–6572.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research 7*, 2541–2563.